1969

Some contributions to the theory of two-phase sampling

Kweku Twum de Graft-Johnson
Iowa State University

SOME CONTRIBUTIONS TO THE THEORY OF TWO-PHASE SAMPLING

by

Kweku Twum de Graft-Johnson

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: Statistics

Approved:

Signature was redacted for privacy.
In Charge of Major Work

Signature was redacted for privacy.
Head of Major Department

Signature was redacted for privacy.
Dean of Graduate College

Iowa State University
Of Science and Technology
Ames, Iowa

1969

TABLE OF CONTENTS

I. INTRODUCTION

A. Definition and Description of Method

With two-phase sampling, a preliminary sample of size n' is selected, and the values of a concomitant variable, X, are determined for all units. Commonly, a subsample of size n is then drawn and the values of the characteristic of interest, Y, are obtained for units in this second phase sample. This second phase sample may, however, be selected independently of the first phase sample. Some statisticians (including Chakravarti et al. [5]) use the term "double sampling" to refer to independent selection of second phase units and reserve the term "two-phase sampling" for dependent second phase sampling. In our presentation, the two terms are used interchangeably. In usual survey practice, an independent second phase sample is generally used only when it is administratively more convenient to do so.

We note that two-phase sampling schemes are applied if the cost of measuring the concomitant variables is small but non-negligible as compared to the cost of measuring the variables of direct interest. Clearly, if the per unit cost of measuring X is negligible, then the best choice is the determination of X for all units in the population.

Two-phase sampling differs from two-stage sampling. In the former, we observe at the two phases sampling units of the same type, but in the latter we consider sampling units of different types at the two stages.

For two-phase sampling, the methods of selection of the preliminary (or "initial") and second phase (or ultimate) samples vary [6, 7]. Two

of the most common ones are:

(a) Sampling without replacement with equal probability in both phases.

(b) Sampling without replacement with equal probability at the first phase but sampling with replacement with arbitrary probabilities at the second phase.

B. Uses of the Technique of Two-phase Sampling

Hereafter, unless otherwise stated, our notation will be as given in Cochran [9], and the sample selection at both phases will be simple random sampling.

The first serious proponent of two-phase sampling was J. Neyman [36]. Its main application has been to improve the precision of estimation [28, 37] through the use of the data obtained at the first phase. Alternatively, the measurements obtained in the initial sample may be utilized to improve the sample design at the second phase. Several examples of the two main purposes of the application of two-phase sampling will now be presented.

Assuming simple random sampling, a common estimator of the population mean, $\bar{Y}$, is:

$$\hat{\bar{Y}} = \frac{\bar{y}}{\bar{x}} \bar{X} \tag{1.1}$$

If $\bar{X}$ is unknown, we can use $\bar{x}'$ to estimate it, where $\bar{x}'$ is the mean of the X variable in a large preliminary simple random sample of size n'. The estimator (1.1) then becomes:

$$\hat{\bar{Y}}_d = \frac{\bar{y}}{\bar{x}}\,\bar{x}' \qquad (1.2)$$

A similar approach can be used in the case of other ratio and regression estimators in both stratified and unstratified as well as single-stage or multi-stage sampling designs. Ratio and regression estimators will be examined in more detail in Sections C and D of this chapter. Some extensions of the existing theory will be presented in Chapters II and III.

As mentioned in the previous paragraph, multi-stage sampling can be combined with two-phase sampling to provide more reliable estimates. For example, in the estimation of the yields of cocoa in Ghana, a two-stage sample of localities and farms can be used for the estimation of areas under cocoa cultivation and a subsample of farms for the estimation of the yields. Thus we have a combination of a two-stage and a two-phase sample design. Examples of this type are given in [55, 59].

Another way in which the technique of two-phase sampling can be utilized to improve the sample design is in the construction of new sampling units [18]. These new units could be made up of any combination of the units selected in the first phase. The new sampling units so constructed can then form the basis for single or multi-stage sampling with arbitrary probabilities of selection. In particular, the first phase units could be selected by simple random sampling and the second phase sample selected with probabilities proportional to the "sizes" (p. p. s.) of the first phase units.

D. Singh and B. D. Singh [51] have considered estimation of the population mean under two different sampling schemes which make use of the preliminary sample for p. p. s. selection at the second stage. For both schemes, they give an unbiased estimator of the population mean, its variance and an estimator of the variance. With both schemes, the preliminary sample is selected by simple random sampling. For the first scheme, the second phase sample is selected with replacement and with probabilities proportional to the values of X in the first phase sample. Under the second scheme, the second phase units are selected by the Rao-Hartley-Cochran scheme [45].

A further use of two-phase sampling is in estimating stratum weights. Given a preliminary sample of size n' and pre-specified boundaries for strata, w_h is an unbiased estimator of W_h where

$$w_h = \frac{n_h'}{n'} = \text{the proportion of elements in the preliminary sample belonging to stratum } h$$

$$W_h = \frac{N_h}{N} = \text{the proportion of the total population of size } N \text{ belonging to stratum } h .$$

These estimators of the population weights can then be used to form an estimator of the population mean

$$\bar{y}_{st} = \sum_{h=1}^{L} w_h \bar{y}_h$$

where $\bar{y}_h$ is the sample mean obtained from the second phase units in stratum h .

Another use of the first phase sample is to construct "optimal" stratification boundaries. Then, one might select a simple random sample within each of the strata, and measure Y for all elements. As noted above, in the classical theory of double sampling with stratification, stratum boundaries are assumed to be fixed prior to sampling. Given L and the data from the preliminary sample, the problem then is to find those boundaries which minimize the conditional variance of the estimator of the population mean, assuming the subsequent use of Neyman, proportional or equal sample allocation. However, to assist in solving this problem, we can use the known theory (one-phase sampling) for constructing optimal stratum boundaries.

Most of the known theory assumes for simplicity that stratification will be based on the variable of interest, Y , rather than on the concomitant variable, X . Dalenius and Hodges [11] have shown that to obtain optimal stratification boundaries, the boundary points must satisfy a certain difference equation. No exact solution to this equation has been found but numerous approximate solutions have been proposed [2, 10, 11, 12, 14, 15]. Cochran [8] has reviewed this literature, and has contrasted the five main methods suggested. Only two of these methods appear satisfactory, namely the "cum$\sqrt{f}$" method [12] and the Ekman approach [14]. Both Herlekar [26, 27] and Serfling [50] have mentioned the inherent logical difficulties in the application of those approximate methods which assume the probability density function (p. d. f.) of the

variable of interest. The more logical and practical course would appear to lie in applying, for example, the cum$\sqrt{f}$ method to the concomitant variable X. It has been conjectured and shown in a few special cases that if intra-stratum correlations are sufficiently high, the optimal stratum boundaries (O.S.B.) for X are also, approximately, the O.S.B. for Y[9, 27].

In Herlekar's first paper [26], he suggests a method whereby given the p.d.f. $f(y, \theta)$ with θ unknown, the preliminary random sample could be used to estimate θ, and thus determine the O.S.B. for Y. The main steps of his procedure are:

(a) Obtain $\hat{\theta}_1$, an estimate of θ, from a preliminary random sample of size $n_1 (< n)$. (θ is a location or scale parameter.)

(b) Using existing theory (e.g. cum$\sqrt{f}$ method), determine the O.S.B. for Y, where $f \doteq f(y, \hat{\theta}_1)$.

(c) Using the O.S.B., choose an independent, stratified random sample of size $n-n_1$. Thus, obtain a second estimate of θ, $\hat{\theta}_2$.

(d) Finally, a linear combination of $\hat{\theta}_1$ and $\hat{\theta}_2$ is used to estimate θ.

Herlekar's second paper [27] treats the case where the stratification variable is the concomitant variable. He investigates the special case where (X, Y) has a bivariate normal distribution, and considers the conditions under which the O.S.B. for X are also the O.S.B. for Y.

Assuming the use of the "cum$\sqrt{f}$" method, Serfling [50] has derived a simple approximation for the minimum (corresponding to the O.S.B.) value of $V(\bar{y}_{st})$. His results will be considered in greater detail in Chapter VI.

Another important area in which two-phase sampling can be used is in the field of analytical surveys [49]. In such surveys, the primary objective is to compare several sectors of a finite population. The following example will illustrate the point: the population of interest may be divisible into four subpopulations represented by a 2 x 2 contigency table i.e. two factors at two levels. The principal objective of the survey may be to compare the two levels of each factor with respect to some specified characteristic. For such a survey, it would be desirable to sample each of the 2 x 2 cells independently. However, it may not be possible to identify prior to sampling the subpopulations to which each individual belongs. In this case, it may be convenient to select a large preliminary sample and identify the subpopulation to which each individual in this preliminary sample belongs. Then, within each cell, independent subsampling can be carried out in accordance with a prescribed sampling rule. Such a sampling rule may maximize the precision of specified estimated comparisons, conditional on the results of the first phase sample.

C. Additional Applications - Ratio Estimators

Before considering the use of double sampling with ratio estimation we list below, for reference purposes, a few ratio-type estimators used with single phase simple random sampling. Some of the properties of these estimators of $R = \bar{Y}/\bar{X}$ are given in the papers referenced. For example, valid asymptotic expressions for the biases and variances of r_1, r_{5B}, r_6 and r_7 and exact expressions for r_2 and r_3 are available in this literature. The estimators are:

(1) the classical ratio estimator [58]

$$r_1 = \bar{y}/\bar{x} \tag{1.3}$$

(2) the mean of ratios estimator

$$r_2 = \frac{1}{n} \sum_{i=1}^{n} (y_i/x_i) \tag{1.4}$$

(3) the Hartley-Ross estimator [25]

$$r_3 = \bar{r} + \frac{n(N-1)}{N(n-1)\bar{X}} (\bar{y} - \bar{r}\bar{x}) \tag{1.5}$$

Recall that r_3 is an unbiased estimator of R.

(4) Pascual's [37]

$$r_4 = \frac{1}{n-1} [n r_1 - r_2] \tag{1.6}$$

(The above estimator, attributed to H. O. Hartley, was first discussed by Pascual [37]).

(5) Quenouille's [13, 41]

$$r_{5A} = \frac{1}{g} \sum_{i=1}^{g} r_{Qi} \tag{1.7a}$$

where the sample of size n is divided at random into g group (each of size m), $r_{Qi} = g\,r_1 - (g-1)r_i'$ and r_i' is the classical ratio estimator calculated from the sample after omitting the units in the ith group. A special case of (1.7a) is obtained by taking $g = 2$:

$$r_{5B} = \frac{1}{2} \sum_{i=1}^{2} r_{Qi} \tag{1.7b}$$

In Chapter II, we shall be considering (1.7b) rather than (1.7a).

(6) Beale (in Tin's [58]).

$$r_6 = r_1 \frac{\{1 + (\frac{1}{n} - \frac{1}{N}) \frac{s_{xy}}{\bar{x}\bar{y}}\}}{\{1 + (\frac{1}{n} - \frac{1}{N}) \frac{s_x^2}{\bar{x}^2}\}} \tag{1.8}$$

(7) Tin's [58]

$$r_7 = r_1[1 + (\frac{1}{n} - \frac{1}{N})(\frac{s_{xy}}{\bar{x}\bar{y}} - \frac{s_x^2}{\bar{x}^2})] \tag{1.9}$$

(8) Mickey's [34, 42]

$$r_8 = \bar{r}_g + \frac{g}{\bar{X}} (\bar{y} - \bar{r}_g \bar{x}) \tag{1.10}$$

where the sample of size n is divided at random into g groups (each of size m), $\bar{r}_g = \sum_{j=1}^{g} r_j'/g$, and r_j' is the classical ratio estimator computed for the sample after omitting the units in the jth group. Recall that r_8 is an unbiased estimator of R.

(9) Tin's (second) [40, 58]

$$r_9 = \frac{g}{g-1} r_1 - \frac{1}{g(g-1)} \sum_{j=1}^{g} \frac{\bar{y}_j}{\bar{x}_j} , \tag{1.11}$$

where $\bar{y}_j$ and $\bar{x}_j$ are the means for group j, the sample being divided at random into g groups, each of size m.

(10) Pascual's (second) [37, 40]

$$r_{10} = r_1 + \frac{1}{(n-1)\bar{X}} (\bar{y} - \bar{r}\bar{x}) \tag{1.12}$$

If we put $\bar{X} = \bar{x}$ in the denominator of the second term of (1.12), $r_4 \doteq r_{10}$.

In a different category is Srivastava's estimator [53]

$$\bar{y}_a = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right)^a \tag{1.13}$$

For $a = 1$, one obtains the product estimator; for $a = 0$, the mean per unit estimator; and for, $a = -1$, the classical ratio estimator.

In addition, we may note that, retaining terms of $O(\frac{1}{n})$, the preferred estimator is as follows:

Range of ρ	Preferred Estimator
$\rho < -\frac{1}{2}\frac{C_x}{C_y}$	product
$-\frac{1}{2}\frac{C_x}{C_y} \leq \rho \leq \frac{1}{2}\frac{C_x}{C_y}$	mean per unit
$\rho > \frac{1}{2}\frac{C_x}{C_y}$	ratio .

Recall that for all the estimators considered above we have assumed that simple random sampling is used. However, if either Lahiri's [33] or Midzuno-Sen's method of selection [48] is employed, it is easily shown (Des Raj [40]) that $E(r_1) = R$. The variance of r_1 under this system is, to $O(\frac{1}{n^2})$,

$$V(r_1) = R^2[\frac{\theta_1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{\theta_2}{n^2}(2C_{21} - C_{30} - C_{12}) + \frac{\theta_3}{n^2}\{3(C_{20} - C_{11})^2 + (1-\rho^2)C_{20}C_{02}\}] \qquad (1.14)$$

where

$$C_{rs} = \frac{E(x_i - \bar{X})^r (y_i - \bar{Y})^s}{\bar{X}^r \bar{Y}^s}$$

$$\theta_1 = \frac{N-n}{N-1}$$

$$\theta_2 = \frac{(N-n)(N-2n)}{(N-1)(N-2)}$$

$$\theta_3 = \frac{N(N-n)(N-n-1)}{(N-1)(N-2)(N-3)}$$

We note that, to $O(\frac{1}{n})$, this result is the same as that of r_1 under simple random sampling.

When trying to compare the various ratio-type estimators with respect to bias, variance and mean square error, many difficulties are encountered. These difficulties stem mainly from the absence, in general, of simple, exact expressions for such quantities. One approach is to use "asymptotic" expansions. Tin used this method to compare r_1, r_{5B}, r_6 and r_7.

Some investigators have employed models to simplify comparisons and to facilitate the interpretation of results. The two models suggested by Durbin [13] and utilized in many subsequent investigations [41, 46]

are:

(a) $$y_i = \alpha + \beta x_i + u_i \qquad (1.15)$$

where $E(u_i | x_i) = 0$, $V(u_i | x_i) = n\delta$, δ is a constant of $O(\frac{1}{n})$, and the x_i's are assumed to be independent, normally distributed random variables each with mean 1 and variance h;

(b) same as (a) with the modification that x_i is distributed as a gamma random variable with parameter m. Rao [42] used both models to compare r_3, r_8, r_9 and r_{10}. Rao and Webster [46] also used the two models to compare r_{5A} (with g = n) with r_1.

A third method of comparing the various ratio-type estimators is to use "Monte Carlo" techniques with or without model assumptions. Rao and Beegle [44] have investigated the small sample efficiencies of eight ratio estimators under the assumption that there is a linear regression of Y on X and X is normally distributed. The estimators they considered were r_1, r_3, r_{5A} (with g = n), r_6, ..., r_{10}. Their Monte Carlo study was based on two models:

(1) With the Lauh and Williams' model, X is normally distributed with mean 10 and variance 4, $y_i = 5(x_i + e_i)$ and e_i is normally distributed with mean 0 and variance 1 and is independent of x_i.

(2) For the second model, (X, Y) is a bivariate normal random variable with the relevant population parameters specified. Two measures are used to compare the estimators. The first one is called "concentration" - the proportion of the estimates generated by the Monte Carlo simulation

which fall in some pre-specified neighborhood of R. The second measure is the interquartile range (as applied to the sample of estimates generated by the Monte Carlo procedure).

Frauendorfer [16] and Rao [43] have carried out similar Monte Carlo investigations, using several actual finite populations.

Of the three main methods (outlined above) for comparison of the estimators, asymptotic expansions are, perhaps, the most general. Analytical comparisons based on model assumptions or comparisons from Monte Carlo studies are not universally valid since they involve rather specific assumptions about the populations under investigation. Of course, the results can be generalized to populations "similar" to those studied. For a general theoretical treatment, it is usually better to make analytical comparisons using a model rather than to use a Monte Carlo - numerical investigation for such comparisons. Of course, if one anticipates repeated sampling of one type of finite population, "Monte Carlo" comparisons based on a "typical" finite population may be preferable (if no adequate model can be postulated).

The ratio estimators defined earlier estimate $R = \bar{Y}/\bar{X}$, and the corresponding estimator of $\bar{Y}$ is obtained by multiplying the estimator of R by $\bar{X}$. If $\bar{X}$ is unknown, it may be estimated from the observations obtained at the first phase of a two-phase sample.

$$\hat{\bar{Y}} = \hat{R}\bar{x}' \qquad (1.16)$$

provides an estimator of $\bar{Y}$, where $\hat{R}$ denotes any estimator of R and $\bar{x}'$ is the estimator of $\bar{X}$ from the preliminary sample. Assuming simple random sampling at both phases, B. V. Sukhatme [54] has contrasted three ratio-type estimators. Assuming that the second phase sample is a subsample of the first phase sample, he compares

$$\bar{y}_{1s} = \frac{\bar{y}}{\bar{x}} \bar{x}' \quad , \tag{1.17}$$

$$\bar{y}_{2s} = \bar{r}\, \bar{x}' \quad , \tag{1.18}$$

and,

$$\bar{y}_{3s} = \bar{r}\, \bar{x}' + \frac{n(n'-1)}{n'(n-1)} (\bar{y} - \bar{r}\, \bar{x}) \quad . \tag{1.19}$$

We note that $\bar{y}_{3s}$ is the two-phase analogue of the Hartley-Ross estimator. Using $\bar{y}_{3s}$ as the standard, Sukhatme investigated the relative efficiencies of $\bar{y}_{1s}$ and $\bar{y}_{2s}$. Writing $\Delta x = x - \bar{X}$, $\Delta r = r - \bar{R}$, where $\bar{X} = E(X)$ and $\bar{R} = E(Y/X)$, he concludes that:

(a) If $$\frac{E[(\Delta r)^2 \Delta x]}{V[\Delta r\, \Delta x]} < -\frac{1}{2\bar{X}}$$

and

and $$\frac{E[(\Delta r)^2 \Delta x]}{V[\Delta r\ \Delta x]} < -\frac{1}{2\bar{R}}$$

then $V(\bar{y}_{3s}) < V(\bar{y}_{2s})$.

(b) Similarly, if

$$\frac{E[(\Delta x)^2 \Delta r]}{V[\Delta r\ \Delta x]} < 0 \qquad (1.20)$$

and $$0 < R < \bar{R} < \sigma_{XY}/\sigma_X^2 \qquad (1.21)$$

then $V(\bar{y}_{3s}) < V(\bar{y}_{1s})$.

In Chapter II (Sections F and G) of this paper, we shall compare $\bar{y}_{3s}$ with $\bar{y}_{1s}$ and other estimators based on r_4 , r_{5B} , r_6 and r_7 , under two different sampling schemes. Biases and variances will be given to $O(\frac{1}{(n')^2})$. In particular, we shall show that to $(\frac{1}{n'})$, condition (1.21) alone implies $V(\bar{y}_{3s}) < V(\bar{y}_{1s})$.

We note here that Goswami and Sukhatme [21] have considered the extension of ratio-type estimation to be multiphase case where several auxiliary variables are used.

D. Additional Applications - Regression Estimators

The regression method of estimation is often recommended when the regression of Y on X is linear but does not pass through the origin. Assuming single-phase simple random sampling, the difference or regression estimator is

$$\bar{y}_R = \bar{y} + \theta(\bar{X} - \bar{x}) \quad . \tag{1.22}$$

When θ is a constant independent of the sample values of (X, Y), (1.22) is referred to as a difference estimator. The value of θ which minimizes $V(\bar{y}_R)$ is easily shown to be $\theta = \beta = S_{XY}/S_X^2$, where β is the population regression coefficient. To estimate β, one generally uses

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \quad .$$

Fuller and Johnson [17] have derived the bias and MSE of $\bar{y}_R$ with $\theta = b$ to $O(\frac{1}{n^2})$.

As in the case of ratio estimation, $\bar{X}$ may be unknown and a large sample estimate of it, $\bar{x}'$ say, may be obtained from the first phase sample of a double sampling procedure. The expression (1.22) then becomes

$$\bar{y}_{Rd} = \bar{y} + \theta(\bar{x}' - \bar{x}) \tag{1.23}$$

We note that

$$E(\bar{y}_{Rd}) = \bar{Y} + Cov(\theta, \bar{x}') - Cov(\theta, \bar{x})$$

If θ is a constant (independent of the sample values), then $\bar{y}_{Rd}$ is an unbiased estimator of $\bar{Y}$. It is again easy to show that $V(\bar{y}_{Rd})$ is minimized if $\theta = \beta$. As in the case of single phase sampling, $\hat{\beta} = b$ is the usual estimator of β. Then

$$\text{bias of } \bar{y}_{Rd} = Cov(b, \bar{x}') - Cov(b, \bar{x}) \qquad (1.24)$$

If the second phase sample is independent of the first phase sample, then (1.24) reduces to:

$$\text{bias of } \bar{y}_{Rd} = - Cov(b, \bar{x}) \qquad (1.25)$$

which is the same as the bias of $\bar{y}_R$. Assuming that the second phase sample is a subsample of the first phase sample, P. V. Sukhatme [56] gives the bias and MSE of $\bar{y}_{Rd}$ to $O(\frac{1}{n'})$. In Chapter III, we shall give these expressions to $O(\frac{1}{(n')^2})$.

E. Comparison of Double Sampling Methods for Ratio and P.P.S. Estimation

Des Raj [38] has considered two methods of double sampling for p. p. s. "estimation" of the population total, Y_T and has contrasted these with double sampling for ratio and regression estimation. In the first case, the preliminary sample of size n' is selected by simple random sampling. The ultimate sample of size n is a subsample of the n' first phase units, selected with p. p. s. with replacement. In this case he concluded that, provided subsampling fractions are small but n is large, double sampling for p. p. s. estimation will be more efficient than double sampling for ratio or regression estimation (under simple random sampling) if, and only if, the p. p. s. estimator is better than the ratio or regression estimator in single phase sampling.

In the second case, the ultimate sample is independent of the first phase sample; that is, the preliminary sample of size n' is used solely to estimate X_T, the population total of the X-variable. The independent second phase sample of size n is selected in accordance with Lahiri's method, which presupposes the existence of an upper bound, M, for the X-variable (sec. 3. 14 of [40]). Again, Des Raj shows that if the second phase sample size is large, the relative efficiency of the p. p. s. "estimator" to the ratio or regression estimator depends solely on the performance of the two types of estimator in the single phase situation.

Some research on the relative efficiencies of estimators under various sampling designs with simple random sampling at the first phase, but with arbitrary probabilities of selection at the second phase has been

done. For example, M. P. Singh [52] has considered three different estimators (corresponding to three second phase designs) of the population total. They have the general form

$$T = \left(\frac{N}{n'}\right)t \text{ , where } t \text{ is an unbiased (second phase) estimator of the first phase total.}$$

To simplify the comparisons, M. P. Singh assumed that the finite population was drawn from an infinite super-population in which X (the concomitant variable) and Y are correlated. The model is

$$y_j = \beta X_j + e_j \quad , \quad j = 1, 2, \ldots, N$$

where

$$E(e_j | X_j) = 0 \, , \quad E(e_j^2 | X_j) = aX_j^g \, , \quad a > 0 \, , \; g > 0$$

and

$$E(e_i e_j | X_i, X_j) = 0 \, .$$

II. SOME PROPERTIES OF VARIOUS RATIO-TYPE ESTIMATORS

A. Sample Design

The sample selection will be in two phases. In the first phase, a sample of n' elements will be selected by a simple random mechanism without replacement from a population of size N, with N large. The second phase sample of size $n (n < n')$ will also be a simple random sample without replacement selected either independently of the first phase sample (Scheme I) or as a subsample of it (Scheme II). This implies that in Scheme I, the f. p. c. at both phases will be ignored. In Scheme II, the first phase f. p. c. will be ignored while the second phase f. p. c. will be retained. However, in one simple case, we shall consider the effect of retaining the f. p. c. at the first phase.

B. Notation

We define a finite population, $U_1, U_2, \ldots U_N$, of N distinguishable elements, N assumed large. On each element U_i, we measure two characteristics (X, Y), where Y is the characteristic of interest and X is the auxiliary or concomitant variable.

We define also the sets $s_{n'}$ and s_n, where

$$s_{n'} = \{U_i: U_i \text{ belongs to the preliminary sample}\}$$

$$s_n = \{U_i: U_i \text{ belongs to the second phase sample}\}.$$

We define further

$$(1) \qquad \bar{x}_n = \frac{1}{n} \sum_{i \epsilon s_n} x_i , \qquad \bar{x}_{n'} = \frac{1}{n'} \sum_{i \epsilon s_{n'}} x_i .$$

(2) $r = Y/X$, $E(r) = \overline{R}$, $R = \overline{Y}/\overline{X}$.

(3) $E(u) = \overline{U}$.

(4) $\Delta u = u - \overline{U}$

$$\delta u = \frac{\Delta u}{\overline{U}}$$

(5) $C_{ij} = E(\delta x)^i(\delta y)^j$. The C_{ij}'s are the so-called product-moment coefficients. It is essential to note that I am defining the C_{ij}'s in terms of the bivariate moments about the mean and not in terms of bivariate cumulants. Hence my notation differs somewhat from that used by Tin [58] but, for $r + s \leq 3$, the bivariate moments about the means, μ_{rs} (defined below), are the same as the bivariate cumulants K_{rs}, used by Tin. A conversion from the μ_{rs} to the K_{ij} notation and vice versa can be performed using expressions given by Kendall and Stuart [29].

(6) $\mu_{ij} = [E(\Delta x)^i(\Delta y)^j]$

$$\sigma_{uv} = E[(\Delta u)(\Delta v)]$$

(7) $$\hat{R} = \frac{\bar{y}_n}{\bar{x}_n}$$

(8) $$s_{uv} = \frac{1}{n-1} \sum_{i \epsilon s_n} (u_i - \bar{u}_n)(v_i - \bar{v}_n)$$

(9) ρ = population coefficient of correlation between X and Y .

ρ_1 = population coefficient of correlation between r and X .

(10) $\beta(\bar{y}_{ij})$ = bias of estimator $\bar{y}_{ij}$.

(11) C_x = co-efficient of variation of X .

C. Assumptions

The following general assumptions will be made:

(1) (a) $X > 0$, $Y > 0$. This assumption, as far as it affects X , is to ensure that the various ratios considered here are well defined. The assumption as far as it refers to Y is for convenience of comparisons only. Under this assumption, R and $\bar{R}$ are both positive.

(b) Unless otherwise stated, $\rho > 0$, and hence $C_{11} > 0$.

(2) Except where otherwise stated, all results will be given to $O(\frac{1}{(n')^2})$

(3) In this chapter, comparisons will be made between different estimators with respect to bias and MSE in the following cases:

(a) The general case in which we make no assumption about the distribution of (x_i, y_i).

(b) The finite population $(y_1, y_2, \ldots y_N)$ is a random sample from an infinite population, the pairs of values (x_i, y_i) satisfying the model:

$$y_i = \alpha + \beta x_i + e_i , \qquad \text{where}$$

$$E(e_i | x_i) = 0 \quad \text{and} \quad E(e_i^2 | x_i) = a x_i^g , \qquad g \geq 0 .$$

(c) Following Durbin [13] and Rao and Webster [46], we assume that (X, Y) satisfy the model:

$$y_i = \alpha + \beta x_i + u_i , \quad \text{where } X \text{ has a gamma distribution with probability density function given by}$$

$$f(x) = \frac{x^{m-1} e^{-x}}{\Gamma(m)} , \qquad m > 0, \; x > 0$$

$$= 0 \qquad \text{otherwise}$$

with $E(u_i | x_i) = 0$ and $E(u_i^2 | x_i) = n\delta$, where $\delta = O(\frac{1}{n})$.

(d) Suppose that there is a quadratic regression of Y on X given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i ,$$

where we assume that X is a normal variable, with the units of measurement chosen in such a way that $E(x_i) = 1$ and $V(x_i) = h$, where $h = O(\frac{1}{n})$. Also, $E(\epsilon | x_i) = 0$, $V(\epsilon | x) = n\delta$, where $\delta = O(\frac{1}{n})$.

The comparisons will be done under two different schemes. In Scheme I, the second phase sample will be independent of the first phase sample while under Scheme II it will be a subsample of it.

D. Basic Results Reproduced without Detailed Proof

The following results attributable to the sources indicated are being reproduced here for ease of reference since they will be used frequently hereafter. They will be stated without proof but with some commentary which in some cases will correspond to a sketch of a proof.

(1) If X and Y are random variables, then Goodman [19] has shown that:

(a) if X and Y are independent,

$$V(XY) = \bar{X}^2 V(Y) + \bar{Y}^2 V(X) + V(X)V(Y) \tag{2.1}$$

(b) if X and Y are not independent,

$$V(XY) = \bar{X}^2 V(Y) + \bar{Y}^2 V(X) + 2\bar{X}\bar{Y}\mu_{11} + 2\bar{X}\mu_{12} + 2\bar{Y}\mu_{21} + \mu_{22} - \mu_{11}^2 \tag{2.2}$$

where

$$\mu_{22} - \mu_{11}^2 = E(\Delta X)^2(\Delta Y)^2 - [E(\Delta X \Delta Y)]^2 = V(\Delta X \Delta Y) \ .$$

Equations (2.1) and (2.2) can easily be verified.

(2) Hansen, Hurwitz and Madow [23] have stated that

$$E(\bar{x}_n - \bar{X})^r(\bar{y}_n - \bar{Y})^s = O\left(\frac{1}{n^{(r+s)/2}}\right) \quad \text{if } r+s \text{ is even}$$

$$= O\left(\frac{1}{n^{(r+s+1)/2}}\right) \quad \text{if } r+s \text{ is odd} \tag{2.3}$$

It is easy to verify result (2.3) in the simple case $r + s \leq 4$ where sampling is with replacement. Hansen et al. have indirectly proved the general result for the case $s = 0$. A proof of the general case is outlined below. Since sampling is without replacement with N large, this is essentially equivalent to sampling being independent from trial to trial. The proof below is based on the multinomial theorem. We shall assume without loss of generality that $\overline{X} = \overline{Y} = 0$.

We note that

$$E(\bar{x}_n^r \bar{y}_n^s) = \frac{1}{n^{r+s}} E[\Sigma x_i]^r [\Sigma y_i]^s . \tag{2.4}$$

The general term of this expression is given by

$$G_t = \frac{f(n, r, s)}{n^{r+s}} E(x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_k^{\alpha_k} y_1^{\beta_1} y_2^{\beta_2} \ldots y_p^{\beta_p}) \tag{2.5}$$

where $\sum_{i=1}^{k} \alpha_i = r$ and $\sum_{j=1}^{p} \beta_j = s$, α_i and β_i taking only zero and positive integral values.

To determine the order of $E(\bar{x}_n^r \bar{y}_n^s)$, we need only consider the highest index of n in the coefficient $f(n, r, s)$ in expression (2.5). Since sampling is independent from trial to trial, we note that G_t vanishes if $\alpha_i = 1$ and the corresponding $\beta_i = 0$ and vice versa. The highest index of n in the coefficient $f(n, r, s)$ is clearly the maximum of $f(n, r, s)$ with respect to the power of n. This index depends on the maximum number of different pairs of factors in (2.5) such that $G_t \neq 0$.

Max. no. of factors such that $G_t \neq 0$ = $\min(r, s) + \frac{\max(r, s) - \min(r, s)}{2}$

if $r + s$ is even.

or = $\min(r, s) + \frac{\max(r, s) - \min(r, s) - 1}{2}$

if $r + s$ is odd.

Hence:

(a) If $r + s$ is even,

max. no. of factors = $\frac{r+s}{2}$ and the different ways of

selecting them = $n(n-1)(n-2) \ldots (n-\frac{r+s}{2}+1)$

= $O(n^{(r+s)/2})$.

From equation (2.4),

$$E(\bar{x}_n^r \bar{y}_n^s) = O(n^{-\frac{r+s}{2}}) \qquad (2.6a)$$

(b) If $r + s$ is odd,

max. no. of factors = $\frac{r+s-1}{2}$

and equation (2.4) leads to

$$E(\bar{x}_n^r \bar{y}_n^s) = O(n^{-\frac{r+s+1}{2}}) . \qquad (2.6b)$$

(3) Fuller and Johnson [17] have shown that given a finite population with mean $\bar{X} \neq 0$, $S_x^2 = \sum_{i=1}^{N} (x_i - \bar{X})^2/N-1$ and $\delta > 0$, there exists a sample size $n_0 \leq N$ such that for $n_0 \leq n \leq N$

$$P\{\left|\frac{\bar{x}_n - \bar{X}}{\bar{X}}\right| \geq 1\} < \delta$$

or alternatively

$$P\{\left|\frac{\bar{x}_n - \bar{X}}{\bar{X}}\right| < 1\} \geq 1 - \delta \ .$$

We can choose $n_0 = \frac{S_x^2}{\delta + S_x^2/N}$ such that for $n_0 \leq n \leq N$

$$P\{\left|\frac{\bar{x}_n - \bar{X}}{\bar{X}}\right| < 1\} \geq 1-\delta$$

and hence the series expansion of

$$[1 + \frac{\bar{x}_n - \bar{X}}{\bar{X}}]^{-1}$$

converges with probability one. Tin [58] has derived a similar result in the case where N is so large that the f. p. c. is neglected.

We assume in our subsequent work that

$$\left|\frac{\bar{x}_n - \bar{X}}{\bar{X}}\right| < 1 \ . \qquad (2.7)$$

The above argument suggests that this assumption is valid with high probability for sufficiently large n. For example, if $S_x^2 = 986$, $\delta = 0.01$, then

$$n_0 = 990 \quad \text{when} \quad N = 1,000$$

and

$$n_0 = 49,650 \text{ when} \quad N = 100,000 \ .$$

It is also of some interest to consider the effect of our assumption that terms of the form $(1 + \delta\bar{x}_n)^{-1}$ can be expanded as an infinite series and the relevant expectations taken to $O(\frac{1}{n^2})$.

Consider as an example, the case where the x_i's are from the gamma distribution with probability density function given by:

$$f(x) = \frac{e^{-x} x^{m-1}}{\Gamma(m)} , \qquad m > 0, \ x > 0$$

$$= 0 \qquad \text{otherwise.}$$

In this example, we have: $E(x) = m$, $C_{20} = \frac{1}{m}$ and $C_{30} = \frac{2}{m^2}$. To $O(\frac{1}{n^2})$,

$$E(\frac{1}{\bar{x}_n}) \doteq \frac{1}{m}(1 + \frac{1}{mn} + \frac{1}{m^2 n^2}) \ . \tag{2.8}$$

Denote this value by $E_a(\frac{1}{\bar{x}_n})$. But Σx_i is distributed as a gamma variable with parameter nm. Hence, the exact value of

$$E(\frac{1}{\bar{x}_n}) = \frac{n}{nm-1} \tag{2.9}$$

From equations (2.8) and (2.9), the proportionate error in using $E_a(\frac{1}{\bar{x}_n})$ is

$$\text{Error} = \frac{\left| E(\frac{1}{\bar{x}_n}) - E_a(\frac{1}{\bar{x}_n}) \right|}{E(\frac{1}{\bar{x}_n})}$$

$$= \frac{1}{n^3 m^3} \tag{2.10}$$

Similarly if we had evaluated $E_a(\frac{1}{\bar{x}_n})$ to $O(\frac{1}{n})$, the relative error would have been $\frac{1}{n^2 m^2}$. Hence, in this example, the proportionate error in using Taylor's approximation instead of the exact value is a function of both n and the parameter m. The approximation is therefore useful only if the product nm is large.

The above arguments in respect of $\delta\bar{x}_n$ can easily be extended to expressions like δs_x^2. Noting that $V(s_x^2) = O(\frac{1}{n})$, we can use the same stochastic argument to motivate the assumption that $|\delta s_x^2| < 1$.

E. Required Preliminary Results

The following results which will be required in Chapters II to IV are stated here for convenience. The first four results are deducible from Tin's work [58] but were independently derived in this presentation. As previously indicated, all results are correct to $O(\frac{1}{n^2})$.

$$E\left(\frac{\bar{y}_n}{\bar{x}_n}\right) = R\left[1+\frac{1}{n}(C_{20}-C_{11})+\frac{1}{n^2}(C_{21}-C_{30}-3C_{20}C_{11}+3C_{20}^2)\right] \quad (2.11)$$

This result was also obtained by Koop, using a different approach [32]

$$V\left(\frac{\bar{y}_n}{\bar{x}_n}\right) = R^2\left[\frac{1}{n}(C_{20}-2C_{11}+C_{02})\right.$$

$$+\frac{1}{n^2}(8C_{20}^2+5C_{11}^2-16C_{20}C_{11}+3C_{20}C_{02}-2C_{30}$$

$$\left.-2C_{12}+4C_{21})\right] \quad (2.12)$$

$$E\left[\frac{\bar{y}_n}{\bar{x}_n}\left(1+\frac{1}{n}\left(\frac{s_{xy}}{\bar{x}_n\bar{y}_n}-\frac{s_x^2}{\bar{x}_n^2}\right)\right)\right] = R\left[1-\frac{1}{n^2}(2C_{21}-2C_{30}+3C_{20}^2-3C_{20}C_{11})\right] \quad (2.13)$$

$$V\left[\frac{\bar{y}_n}{\bar{x}_n}\left\{1+\frac{1}{n}\left(\frac{s_{xy}}{\bar{x}_n\bar{y}_n}-\frac{s_x^2}{\bar{x}_n^2}\right)\right\}\right] = R^2\left[\frac{1}{n}(C_{20}-2C_{11}+C_{02})\right.$$

$$\left.+\frac{1}{n^2}(2C_{20}^2+C_{11}^2-4C_{20}C_{11}+C_{20}C_{02})\right] \quad (2.14)$$

The following results are also given to $O(\frac{1}{n^2})$:

$$E[\delta^2\bar{x}_n\delta^2\bar{y}_n] = \frac{1}{n^2}[2C_{11}^2+C_{20}C_{02}],$$

$$E[\delta^4 \bar{x}_n] = \frac{3C_{20}^2}{n^2}$$

$$E[(\delta \bar{x}_n)^3 (\delta \bar{y}_n)] = \frac{3}{n^2} C_{20} C_{11}$$

$$E(\delta \bar{x}_n \delta s_{xy}) = \frac{1}{n} \frac{C_{21}}{C_{11}}$$

$$E(\delta \bar{x}_n \delta s_x^2) = \frac{1}{n} \frac{C_{30}}{C_{20}}$$

$$E(\delta \bar{y}_n \delta s_{xy}) = \frac{1}{n} \frac{C_{12}}{C_{11}}$$

$$E(\delta \bar{y}_n \delta \bar{x}_n \delta s_{xy}) = \frac{1}{n^2} \left[\frac{C_{22}}{C_{11}} - \frac{C_{20} C_{02}}{C_{11}} - 2C_{11}\right]$$

$$E(\delta s_x^2 \delta s_{xy} \delta \bar{x}_n) = \frac{1}{n^2} \left[\frac{C_{41}}{C_{20} C_{11}} - \frac{4C_{21}}{C_{11}} - 2 \frac{C_{30}}{C_{20}}\right]$$

$$E(\delta s_x^2 \delta s_{xy} \delta \bar{x}_n \delta \bar{y}_n) = \frac{1}{n^2} \left[\frac{C_{31}}{C_{20}} + \frac{C_{21}^2}{C_{20} C_{11}} + \frac{C_{30} C_{12}}{C_{20} C_{11}} - C_{11}\right]$$

$$E(\delta^2 s_x^2 \delta \bar{x}_n) = \frac{1}{n^2} \left[\frac{C_{50}}{C_{20}^2} - 6 \frac{C_{30}}{C_{20}}\right]$$

$$E(\delta^2 s_x^2 \delta s_{xy} \delta \bar{x}_n) = \frac{1}{n^2} \left[\frac{C_{40} C_{21}}{C_{20}^2 C_{11}} + \frac{2C_{31} C_{30}}{C_{20}^2 C_{11}} - \frac{C_{21}}{C_{11}} - 2 \frac{C_{30}}{C_{20}}\right]$$

$$E(\delta^2 s_x^2 \delta\bar{y}_n \delta\bar{x}_n) = \frac{1}{n^2}\left[\frac{C_{40}C_{11}}{C_{20}^2} + \frac{2C_{30}C_{21}}{C_{20}^2} - C_{11}\right]$$

$$E(\delta^3 s_x^2 \delta\bar{x}_n) = \frac{3}{n^2}\left[\frac{C_{40}C_{30}}{C_{20}^3} - \frac{C_{30}}{C_{20}}\right]$$

$$E(\delta s_x^2 \delta\bar{x}_n \delta\bar{y}_n) = \frac{1}{n^2}\left[\frac{C_{31}}{C_{20}} - 3C_{11}\right]$$

$$E(\delta^2 s_x^2) = \frac{1}{n}\left[\frac{C_{40}}{C_{20}^2} - 1 + \frac{2}{n}\right]$$

$$E(\delta s_x^2 \delta s_{xy}) = \frac{1}{n}\left[\frac{C_{31}}{C_{20}C_{11}} - 1 + \frac{2}{n}\right]$$

$$E(\delta^2 s_{xy}) = \frac{1}{n}\left[\frac{C_{22}}{C_{11}^2} + \frac{C_{20}C_{02}}{nC_{11}^2} + \frac{1}{n} - 1\right]$$

$$E(\delta s_{xy} \delta^2\bar{x}_n) = \frac{1}{n^2}\left[\frac{C_{31}}{C_{11}} - 3C_{20}\right]$$

$$E(\delta^2 s_{xy} \delta^2\bar{x}_n) = \frac{1}{n^2}\left[\frac{C_{20}C_{22}}{C_{11}^2} + \frac{2C_{21}^2}{C_{11}^2} - C_{20}\right]$$

$$E(\delta s_x^2 \delta^2\bar{x}_n) = \frac{1}{n^2}\left[\frac{C_{40}}{C_{20}} - 3C_{20}\right]$$

$$E(\delta s_{xy}\delta s_x^2\delta^2\bar{x}_n) = \frac{1}{n^2}\left[\frac{C_{31}}{C_{11}} + \frac{2C_{30}C_{21}}{C_{11}C_{20}} - C_{20}\right]$$

$$E(\delta^2 s_x^2\delta^2\bar{x}_n) = \frac{1}{n^2}\left[\frac{C_{40}}{C_{20}} + \frac{2C_{30}^2}{C_{20}^2} - C_{20}\right]$$

$$E(\delta^2\bar{x}_n\delta\bar{y}_n\delta s_x^2) = \frac{1}{n^2}\left[C_{21} + \frac{2C_{30}C_{11}}{C_{20}}\right]$$

$$E(\delta\bar{x}_n\delta^2\bar{y}_n\delta s_x^2) = \frac{1}{n^2}\left[\frac{C_{30}C_{02}}{C_{20}} + \frac{2C_{11}C_{21}}{C_{20}}\right]$$

$$E(\delta^2\bar{x}_n\delta\bar{y}_n\delta s_{xy}) = \frac{1}{n^2}\left[2C_{21} + \frac{C_{20}C_{12}}{C_{11}}\right]$$

$$E(\delta\bar{x}_n\delta^2\bar{y}_n\delta s_{xy}) = \frac{1}{n^2}\left[2C_{12} + \frac{C_{02}C_{21}}{C_{11}}\right]$$

$$E(\delta s_x^2\delta^3\bar{x}_n) = \frac{3C_{30}}{n^2}$$

$$E(\delta s_{xy}\delta^3\bar{x}_n) = \frac{3C_{21}}{n^2}$$

$$E(\delta^2\bar{y}_n\delta s_x^2) = \frac{1}{n^2}\left[\frac{C_{22}}{C_{20}} - \frac{2C_{11}^2}{C_{20}} - C_{02}\right]$$

$$E(\delta^2 s_x^2 \delta\bar{y}_n) = \frac{1}{n^2}\left[\frac{C_{41}}{C_{20}^2} - \frac{2C_{21}}{C_{20}} - \frac{4C_{30}C_{11}}{C_{20}^2}\right]$$

$$E(\delta^3 s_x^2) = \frac{1}{n^2}\left[\frac{C_{60}}{C_{20}^3} - \frac{3C_{40}}{C_{20}^2} - \frac{6C_{30}^2}{C_{20}^3} + 2\right]$$

$$E(\delta\bar{y}_n \delta s_x^2 \delta s_{xy}) = \frac{1}{n^2}\left[\frac{C_{32}}{C_{20}C_{11}} - \frac{4C_{21}}{C_{20}} - \frac{C_{12}}{C_{11}} - \frac{C_{30}C_{02}}{C_{20}C_{11}}\right]$$

$$E(\delta^2 s_x^2 \delta s_{xy}) = \frac{1}{n^2}\left[\frac{C_{51}}{C_{20}^2 C_{11}} - \frac{6C_{30}C_{21}}{C_{20}^2 C_{11}} - \frac{2C_{31}}{C_{20}C_{11}} - \frac{C_{40}}{C_{20}^2} + 2\right]$$

$$E(\delta^3 s_{xy}) = \frac{1}{n^2}\left[\frac{C_{33}}{C_{11}^3} - \frac{3C_{22}}{C_{11}^2} - \frac{6C_{21}C_{12}}{C_{11}^3} + 2\right] \qquad (2.15)$$

We recall that sampling at either phase is simple random sampling without replacement, with N large. In the case of the second phase sample being a subsample of the first phase sample

$$E[(\Delta\bar{x}_{n'}) | x_1, x_2, \ldots x_n, y_1, y_2, \ldots, y_n] = \frac{n(N-n')}{n'(N-n)} \Delta\bar{x}_n$$

$$= \frac{n}{n'} \Delta\bar{x}_n \qquad (2.16)$$

for infinitely large N .

It can also be easily verified that

$$E[\Delta^2\bar{x}_{n'}) \,|\, x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n]$$

$$= [\frac{n(N-n')}{n'(N-n)}]^2 \Delta^2\bar{x}_n + \frac{(N-n')(n'-n)}{(n')^2(N-n)(N-n-1)} \{(N-1)S_X^2$$

$$- (n-1)s_x^2 - \frac{nN}{N-n}\Delta^2\bar{x}_n\}$$

$$= \frac{n^2}{(n')^2} \Delta^2\bar{x}_n + (\frac{1}{n'} - \frac{n}{(n')^2})\mu_{20} \qquad (2.17)$$

for large N.

It should be noted that for the results above we have assumed that N is so large that sampling is independent from trial to trial. Some of these expressions are rather complicated in the case of sampling without replacement with N not very large. For example, Hansen et al. [23] give the equivalent expression for $E(\Delta s_x^2)$ in the case of sampling without replacement (see 2.39d). The expression does not lend itself to the type of further manipulation we need to do in this paper.

F. Scheme I

In this section, we shall compare the biases and MSE's of six estimators under Scheme I where the second phase sample is selected independently of the first phase sample. The estimators to be considered are:

(1) the classical ratio estimator

$$\bar{y}_{11} = \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_n', \tag{2.18}$$

(2) the Hartley-Ross estimator

$$\bar{y}_{12} = \bar{r}_n \bar{x}_n' + \frac{n}{n-1}(\bar{y}_n - \bar{r}_n \bar{x}_n) \tag{2.19}$$

(3) Pascual's estimator

$$\bar{y}_{13} = [\frac{n}{n-1} \frac{\bar{y}_n}{\bar{x}_n} - \frac{1}{n-1} \bar{r}_n]\bar{x}_n', \tag{2.20}$$

(4) Beale's estimator

$$\bar{y}_{14} = \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_n', [\frac{1 + \frac{1}{n} \frac{s_{xy}}{\bar{x}_n \bar{y}_n}}{1 + \frac{1}{n} \frac{s_x^2}{\bar{x}_n^2}}] \tag{2.21}$$

(5) Tin's estimator

$$\bar{y}_{15} = \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_n', [1 + \frac{1}{n}(\frac{s_{xy}}{\bar{x}_n \bar{y}_n} - \frac{s_x^2}{\bar{x}_n^2})] \tag{2.22}$$

(6) Quenouille's estimator

$$\bar{y}_{16} = [2\hat{R} - \frac{1}{2}(\hat{R}_1 + \hat{R}_2)]\bar{x}_n, \tag{2.23}$$

where the second-phase sample of size n is divided at random into two groups of equal size and $\hat{R}$ is the ordinary ratio estimator calculated from all units of the second-phase sample while $\hat{R}_i (i = 1, 2)$ is the ratio estimator for the ith random group.

From (2.11),

$$E(\bar{y}_{11}) = \bar{Y}[1+\frac{1}{n}(C_{20}-C_{11})+\frac{1}{n^2}(C_{21}-C_{30}-3C_{20}C_{11}+3C_{20}^2)] \tag{2.24}$$

We recall from elementary sampling theory e. g. Cochran [9] that

$$E(\bar{y}_{12}) = \bar{Y} \tag{2.25}$$

Using equation (2.24), it is also easy to show that

$$E(\bar{y}_{13}) = \bar{Y}[1+\frac{1}{n-1}(C_{20}-C_{11})+\frac{1}{n(n-1)}(C_{21}-C_{30}-3C_{20}C_{11}+3C_{20}^2)$$

$$+\frac{1}{n-1}\frac{\sigma_{rX}}{\bar{Y}}]. \tag{2.26a}$$

Equation (2.26a) can be simplified by substituting $\bar{Y}-\overline{RX}$ for σ_{rX} and hence

$$E(\bar{y}_{13}) = \bar{Y}[1 + \frac{1}{n}(C_{20} - C_{11}) + \frac{1}{n}(1 - \frac{\bar{R}}{R})$$

$$+ \frac{1}{n^2}(C_{20} - C_{11} + C_{21} - C_{30} - 3C_{20}C_{11} + 3C_{20}^2$$

$$+ 1 - \frac{\bar{R}}{R})] . \qquad (2.26b)$$

The bias of $\bar{y}_{13}$ is in general of $O(\frac{1}{n})$ but in certain cases can be shown to be of $O(\frac{1}{n^2})$. It has been shown by W. H. Williams [62] that

$$\mathrm{Cov}(\frac{y}{u}, \frac{x}{v}) \doteq \frac{1}{\bar{U}\bar{V}} \mathrm{Cov}(y - \frac{\bar{V}}{\bar{U}} u , x - \frac{\bar{X}}{\bar{V}} v) \qquad (2.27)$$

This is just the product of the two first order Taylor series expansions of $\frac{y}{u} - E(\frac{y}{u})$ and $\frac{x}{v} - E(\frac{x}{v})$. If we assume that this approximation is reasonable for $\mathrm{Cov}(\frac{y}{x}, \frac{x}{1})$, then we have:

$$\sigma_{rx} \doteq \frac{1}{\bar{X}} \mathrm{Cov}(y - Rx , x - \bar{X})$$

$$= \bar{Y}(C_{11} - C_{20}) . \qquad (2.28)$$

Approximation (2.28) can easily be justified if $|\delta x| << 1$ [20] or, what is roughly equivalent, if the product-moment coefficients C_{rs} are negligible for $r + s > 2$.

Using equation (2.28), equation (2.26b) reduces to:

$$E(\bar{y}_{13}) \doteq \bar{Y}\left[1+\frac{1}{n^2}\{3C_{20}(C_{20}-C_{11})+C_{21}-C_{30}\}\right]. \quad (2.29)$$

Next, we have:

$$E(\bar{y}_{14}) = \bar{Y}\left[1-\frac{2}{n^2}\{(C_{21}-C_{30})+C_{20}(C_{20}-C_{11})\}\right]. \quad (2.30)$$

$$E(\bar{y}_{15}) = \bar{Y}\left[1-\frac{1}{n^2}\{2(C_{21}-C_{30})+3C_{20}(C_{20}-C_{11})\}\right]. \quad (2.31)$$

Finally

$$E(\bar{y}_{16}) = \bar{Y}\left[1-\frac{2}{n^2}\{(C_{21}-C_{30})+3C_{20}(C_{20}-C_{11})\}\right]. \quad (2.32)$$

The Hartley-Ross estimator, $\bar{y}_{12}$, is the only unbiased estimator among the six considered above. If the second phase sample of size n is large enough so that terms of $O(\frac{1}{n^2})$ are negligible, then $\bar{y}_{14}$, $\bar{y}_{15}$ and $\bar{y}_{16}$ have negligible bias. In this case, only $\bar{y}_{11}$ and $\bar{y}_{13}$ have biases worth considering. It is easy to verify that:

1. $|\beta(\bar{y}_{11})| > |\beta(\bar{y}_{13})|$ if $\rho < \frac{C_x}{C_y}$, $\rho_1 < 0$ and $\left|\frac{\sigma_{rX}}{\bar{Y}}\right| < 2(C_{20}-C_{11})$

or $\rho > \frac{C_x}{C_y}$, $\rho_1 > 0$ and $\frac{\sigma_{rX}}{\bar{Y}} < 2|C_{20}-C_{11}|$.

2. $|\beta(\bar{y}_{11})| < |\beta(\bar{y}_{13})|$ if $\rho > \frac{C_x}{C_y}$ and $\rho_1 < 0$ or $\rho < \frac{C_x}{C_y}$ and $\rho_1 > 0$.

In view of (2.28), the pairs of inequalities in (2) above appear unlikely to occur.

If n is not large enough so that terms of $O(\frac{1}{n^2})$ cannot be neglected, then no general ranking of the five biased estimators with respect to the magnitudes of their biases is possible. However if we consider only $\bar{y}_{14}$, $\bar{y}_{15}$, $\bar{y}_{16}$, these can be ranked if $C_{21} = C_{30}$, as in a bivariate normal population. In this case

$$|\beta(\bar{y}_{14})| \leq |\beta(\bar{y}_{15})| \leq |\beta(\bar{y}_{16})| .$$

We next proceed to find the variances of the various estimators. Applying equation (2.1), we note that

$$V(\bar{y}_{11}) = \bar{X}^2 V(\frac{\bar{y}_n}{\bar{x}_n}) + [E(\frac{\bar{y}_n}{\bar{x}_n})]^2 V(\bar{x}_{n'}) + V(\bar{x}_{n'}) V(\frac{\bar{y}_n}{\bar{x}_n}) .$$

Hence from equations (2.11) and (2.12), we obtain;

$$V(\bar{y}_{11}) = \bar{Y}^2 [\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'} C_{20}$$

$$+ \frac{1}{n^2}(4C_{21} - 2C_{12} - 2C_{30} - 16C_{20}C_{11}$$

$$+ 3C_{20}C_{02} + 8C_{20}^2 + 5C_{11}^2)$$

$$+ \frac{1}{nn'}(3C_{20}^2 - 4C_{20}C_{11} + C_{20}C_{02})] . \qquad (2.33a)$$

From equations (2.24) and (2.33a),

$$MSE(\bar{y}_{11}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}C_{20}$$

$$+ \frac{1}{n^2}(4C_{21} - 2C_{12} - 2C_{30} - 18C_{20}C_{11}$$

$$+ 3C_{20}C_{02} + 9C_{20}^2 + 6C_{11}^2)$$

$$+ \frac{1}{nn'}(3C_{20}^2 - 4C_{20}C_{11} + C_{20}C_{02})] . \qquad (2.33b)$$

To obtain the variance of $\bar{y}_{12}$, we write

$$\bar{y}_{12} = \bar{r}_n \bar{x}_{n'} + \bar{z}_n , \qquad (2.34a)$$

where $$\bar{z}_n = \frac{n}{n-1}(\bar{y}_n - \bar{r}_n \bar{x}_n) .$$

Assuming N large, Goodman and Hartley [20] have obtained the variance of $\bar{z}_n$ as

$$V(\bar{z}_n) = \frac{1}{n}[E\{(\Delta x)^2(\Delta r)^2\} + \frac{\sigma_r^2\sigma_X^2}{n-1} - \frac{n-2}{n-1}\sigma_{rX}^2] . \qquad (2.34b)$$

Again they have shown that

$$Cov(\bar{r}_n, \bar{z}_n) = 2n^{-1} E[(\Delta x)(\Delta r)^2] \qquad (2.34c)$$

But from equation (2.34a),

$$V(\bar{y}_{12}) = V(\bar{r}_n\bar{x}_{n'}) + 2\bar{X}\,Cov(\bar{r}_n, \bar{z}_n) + V(\bar{z}_n) \qquad (2.34d)$$

since $\bar{x}_{n'}$ is independent of both $\bar{r}_n$ and $\bar{z}_n$. Hence combining equations (2.34b) - (2.34d) and recalling from Goodman and Hartley that $\bar{y}_n - \bar{R}\bar{x}_n = \Delta x\,\Delta r + \bar{X}\,\Delta r$, we obtain

$$V(\bar{y}_{12}) = \frac{1}{n}(\sigma_Y^2 + \bar{R}^2\sigma_X^2 - 2\bar{R}\sigma_{XY}) + \frac{1}{n'}\bar{R}^2\sigma_X^2$$

$$+ \frac{1}{n^2}(\sigma_r^2\sigma_X^2 + \sigma_{rX}^2) + \frac{1}{nn'}(\sigma_r^2\sigma_X^2) . \qquad (2.34e)$$

Since $\bar{y}_{12}$ is unbiased,

$$MSE(\bar{y}_{12}) = V(\bar{y}_{12}) . \qquad (2.34f)$$

Next, we consider $MSE(\bar{y}_{13})$. From result (2.1),

$$V(\bar{y}_{13}) = \frac{\sigma_X^2}{n'}[E(r_4)]^2 + \bar{X}^2 V(r_4) + V(\bar{x}_{n'})V(r_4) \qquad (2.35a)$$

where

$$r_4 = [\frac{n}{n-1}\frac{\bar{y}_n}{\bar{x}_n} - \frac{1}{n-1}\bar{r}_n].$$

It is easy to show that to $O(\frac{1}{n})$

$$Cov(\frac{\bar{y}_n}{\bar{x}_n}, \bar{r}_n) = \frac{R}{n}(\frac{\sigma_{rY}}{\bar{Y}} - \frac{\sigma_{rX}}{\bar{X}}) \qquad (2.35b)$$

Hence from equations (2.11), (2.12), (2.35a) and (2.35b), we obtain

$$V(\bar{y}_{13}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}C_{20}$$

$$+ \frac{1}{n^2}(8C_{20}^2 - 16C_{20} + 5C_{11}^2 + 3C_{20}C_{02}$$

$$- 2C_{30} + 4C_{21} - 2C_{12} + 2C_{20} + 2C_{02}$$

$$- 4C_{11} - \frac{2\sigma_{rY}}{R\bar{Y}} + \frac{2\sigma_{rX}}{R\bar{X}})$$

$$+ \frac{1}{nn'}\{C_{20}(3C_{20} - 4C_{11} + C_{02} + 2 - \frac{2\bar{R}}{R})\}]. \quad (2.35c)$$

Hence

$$MSE(\bar{y}_{13}) = \bar{Y}^2[\frac{1}{n}\{C_{02} - 2C_{11} + C_{20}\} + \frac{1}{n'}C_{20}$$

$$+ \frac{1}{n^2}\{9(C_{20} - C_{11})^2 + 3(1-\rho^2)C_{20}C_{02}$$

$$+ 2(2C_{21} - C_{30} - C_{12})$$

$$+ 2(2C_{20} - 3C_{11} + C_{02}) + (1 - \frac{\bar{R}}{R})^2$$

$$- \frac{2\bar{R}}{R}(C_{20} - C_{11}) - \frac{2}{\bar{Y}}(\frac{\sigma_{rY}}{R} - \sigma_{rX})\}$$

$$+ \frac{1}{nn'}\{C_{20}(3C_{20} - 4C_{11} + C_{02} + 2 - \frac{2\bar{R}}{R})\}]. \quad (2.35d)$$

Following the argument leading to (2.28), we can again simplify (2.35d) by making a similar assumption and writing

$$\sigma_{rY} \doteq \frac{\bar{Y}^2}{\bar{X}}(C_{02}-C_{11}) . \qquad (2.35e)$$

Substituting expressions (2.28) and (2.35e) in expression (2.35d), we obtain:

$$MSE(\bar{y}_{13}) \doteq \bar{Y}^2[\frac{1}{n}\{C_{02}-2C_{11}+C_{20}\}+\frac{1}{n'}C_{20}$$

$$+\frac{1}{n^2}\{8(C_{20}-C_{11})^2+3(1-\rho^2)C_{20}C_{02}$$

$$+2(2C_{21}-C_{30}-C_{12})\}$$

$$+\frac{1}{nn'}\{(C_{20}-C_{11})^2+(1-\rho^2)C_{20}C_{02}\}] . \qquad (2.35f)$$

From Tin's paper [58] and (2.1), we obtain

$$V(\bar{y}_{14}) = \bar{Y}^2[\frac{1}{n}(C_{02}-2C_{11}+C_{20})+\frac{1}{n'}C_{20}$$

$$+\frac{1}{n^2}(2C_{20}^2-4C_{20}C_{11}+C_{11}^2+C_{20}C_{02})$$

$$+\frac{1}{nn'}(C_{20}^2+C_{20}C_{02}-2C_{20}C_{11})] . \qquad (2.36a)$$

Since the bias of $\bar{y}_{14}$ is of $O(\frac{1}{n^2})$, we have, to $O(\frac{1}{n^2})$,

$$MSE(\bar{y}_{14}) = V(\bar{y}_{14}) . \tag{2.36b}$$

Similarly, it is easy to show that

$$MSE(\bar{y}_{15}) = V(\bar{y}_{15}) = V(\bar{y}_{14}) . \tag{2.37}$$

Finally, we compute the variance and MSE of $\bar{y}_{16}$.

$$V(\bar{y}_{16}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'} C_{20}$$

$$+ \frac{1}{n^2}(4C_{20}^2 - 8C_{20}C_{11} + 2C_{11}^2 + 2C_{20}C_{02})$$

$$+ \frac{1}{nn'}(C_{20}^2 + C_{20}C_{02} - 2C_{20}C_{11})] . \tag{2.38a}$$

Again, since the bias of $\bar{y}_{16}$ is of $O(\frac{1}{n^2})$,

$$MSE(\bar{y}_{16}) = V(\bar{y}_{16}) , \quad \text{to } O(\frac{1}{(n')^2}) . \tag{2.38b}$$

We note that to $O(\frac{1}{n'})$, all the estimators except $\bar{y}_{12}$ have the same MSE. Thus comparisons based on terms up to $O(\frac{1}{n'})$ will fail to discriminate between any of the estimators. Hence the need to compute MSE's of the six estimators to $O(\frac{1}{(n')^2})$.

We have assumed that sampling is being done from a finite population with N large with respect to both n and n' . The results for sampling from a finite population with the f. p. c. retained are similar to those already derived except that the coefficients of the parameters become more complicated. For example, the expectation, variance and MSE of $\bar{y}_{11}$ for a finite population are as given below. We first define

$$\theta_1 = \frac{N-n}{N-1} , \qquad \theta_1' = \frac{N-n'}{N-1}$$

$$\theta_2 = \frac{(N-n)(N-2n)}{(N-1)(N-2)} ,$$

$$\theta_3 = \frac{N(N-n)(N-n-1)}{(N-1)(N-2)(N-3)} .$$

Using (2. 1) and Sukhatme's results on symmetric functions [56], we obtain:

$$E_G(\bar{y}_{11}) = \bar{Y}[1 + \frac{\theta_1}{n}(C_{20} - C_{11}) + \frac{\theta_2}{n^2}(C_{21} - C_{30})$$

$$+ \frac{3\theta_3}{n^2} C_{20}(C_{20} - C_{11})] \qquad (2.39a)$$

$$V_G(\bar{y}_{11}) = \bar{Y}^2\left[\frac{\theta_1}{n}(C_{20}-2C_{11}+C_{02}) + \frac{\theta_1'}{n'}C_{20}\right.$$

$$+ \frac{1}{n^2}\{-2\theta_2(C_{12}-2C_{21}+C_{30}) - \theta_1^2(C_{20}-C_{11})^2$$

$$+ \theta_3[9(C_{20}-C_{11})^2+3(1-\rho^2)C_{20}C_{02}]\}$$

$$\left.+ \frac{1}{nn'}\{\theta_1\theta_1'C_{20}(3C_{20}-4C_{11}+C_{02})\}\right]. \qquad (2.39b)$$

Finally from (2.39a) and (2.39b),

$$MSE_G(\bar{y}_{11}) = \bar{Y}^2\left[\frac{\theta_1}{n}(C_{20}-2C_{11}+C_{02}) + \frac{\theta_1'}{n'}C_{20}\right.$$

$$+ \frac{1}{n^2}\{-2\theta_2(C_{12}-2C_{21}+C_{30})+\theta_3[9(C_{20}-C_{11})^2$$

$$+ 3(1-\rho^2)C_{20}C_{02}]\}$$

$$\left.+ \frac{1}{nn'}\{\theta_1\theta_1'C_{20}(3C_{20}-4C_{11}+C_{02})\}\right]. \qquad (2.39c)$$

We note that if we put $\theta_i = \theta_i' = 1$ $(i = 1, 2, 3)$, we obtain the results in (2.24), (2.33a) and (2.33b) respectively. We note that our f.p.c. terms

differ somewhat from Tin's. The difference appears to lie in his writing, for example, $V(\bar{x}_n)$ as $\frac{N-n}{N}\frac{\sigma_X^2}{n}$ instead of $\frac{N-n}{N-1}\frac{\sigma_X^2}{n}$. We also recall from Hansen et al. [23] that for the case where the f.p.c. is not ignored,

$$E(\delta^2 s_x^2) = \frac{(N-1)^2}{N^2(n-1)^2}\{\frac{(n-1)^2}{n} - \frac{n-1}{n(N-1)}[(n-2)(n-3)-(n-1)]$$

$$- \frac{4(n-1)(n-2)(n-3)}{n(N-1)(N-2)} - \frac{6(n-1)(n-2)(n-3)}{n(N-1)(N-2)(N-3)}\} \frac{C_{40}}{C_{20}^2}$$

$$+ \frac{(N-1)^2}{N^2(n-1)^2}\{\frac{(n-1)N}{n(N-1)}[(n-1)^2+2]$$

$$+ \frac{2(n-1)(n-2)(n-3)N}{n(N-1)(N-2)} + \frac{3(n-1)(n-2)(n-3)N}{n(N-1)(N-2)(N-3)}$$

$$- \frac{N^2(n-1)^2}{(N-1)^2}\} \quad . \tag{2.39d}$$

As shown in equations (2.15), this reduces to $E(\delta^2 s_x^2) = \frac{1}{n}[\frac{C_{40}}{C_{20}^2} - 1 + \frac{2}{n}]$, where the latter result is correct to $O(\frac{1}{n^2})$. We shall later apply this and some of the other results in (2.15) in deriving relevant MSE's.

We have chosen to consider sampling from a finite population with N sufficiently large (as compared to n and n') instead of the more general case mainly to simplify our computations, as the above examples show.

We now proceed to compare MSE's of relevant pairs of estimators. Under assumption 3(a) of Section C of this chapter, we have, from equations (2.33b) and (2.34f),

$$MSE(\bar{y}_{12}) - MSE(\bar{y}_{11}) = \bar{Y}^2[(\frac{\bar{R}}{R} - 1)\{(\frac{1}{n}+\frac{1}{n'})(\frac{\bar{R}}{R}+1)C_{20} - \frac{2}{n}C_{11}\}$$

$$+ \frac{1}{n^2}\{\frac{\sigma_r^2}{R^2}C_{20} + \frac{\sigma_{rX}^2}{\bar{Y}^2} + 2(C_{12}+C_{30}-2C_{21})$$

$$- 9(C_{20}-C_{11})^2 - 3(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{C_{20}(\frac{\sigma_r^2}{R^2} + 4C_{11} - 3C_{20} - C_{02})\}] \qquad (2.40a)$$

We note that omitting terms of $O(\frac{1}{n^2})$ the above comparison is relevant to all the other estimators; that is, ignoring terms of $O(\frac{1}{n^2})$, the expression $MSE(\bar{y}_{12}) - MSE(\bar{y}_{1j})$ is identical for $j = 1, 3, 4, 5, 6$.

In this case,

$$MSE(\bar{y}_{12}) - MSE(\bar{y}_{11}) \geq 0$$

if either

(1) $\bar{R} \geq R$ and $\rho \leq (\frac{\bar{R}+R}{2R})(1+\frac{n}{n'})\frac{C_x}{C_y}$ (2.40b)

or

(2) $\bar{R} < R$ and $\rho > (\frac{\bar{R}+R}{2R})(1+\frac{n}{n'})\frac{C_x}{C_y}$. (2.40c)

We note that condition $\bar{R} \geq R$ is equivalent to ρ_1, the population coefficient of correlation between r and X, being non-positive. Again, we note, as pointed out by Cochran [9], that if x_i is the value of y_i at some previous time, then $\frac{C_x}{C_y}$ is likely to be equal to 1. Hence if this assumption is valid, condition (2.40b) is likely to be satisfied in every case in which $\rho_1 < 0$, since the second inequality merely reduces to $\rho \leq 1$.

Again, if we apply (2.28), then, to $O(\frac{1}{n'})$,

$$MSE_a(\bar{y}_{12}) - MSE(\bar{y}_{11}) \geq 0 \qquad \text{if}$$

$$\text{either} \quad \text{(i)} \quad \rho_1 \leq 0 \quad \text{and} \quad \rho < \frac{\frac{C_x}{C_y}(1+\frac{n}{n'})(2+C_x^2)}{2 + C_x^2(1+\frac{n}{n'})} \tag{2.40d}$$

$$\text{or} \quad \text{(ii)} \quad \rho_1 > 0 \quad \text{and} \quad \rho > \frac{\frac{C_x}{C_y}(1+\frac{n}{n'})(2+C_x^2)}{2 + C_x^2(1+\frac{n}{n'})}. \tag{2.40e}$$

As in the case of (2.40b), the second inequality in (2.40d) simplifies considerably in the special case where $\frac{C_x}{C_y} \doteq 1$ and n' is so large that $(1 + \frac{n}{n'}) \doteq 1$. Then the second inequality in (2.40d) reduces to $\rho \leq 1$, which again is satisfied in every case.

We could have guessed that to $O(\frac{1}{n'})$ the Hartley-Ross estimator, $\bar{y}_{12}$, would be inferior to $\bar{y}_{11}$ since the bias terms are not included

for $\bar{y}_{11}$ to this order and the Hartley-Ross estimator is also generally inferior to the classical ratio estimator in the single phase case.

Applying (2.28) to the whole of (2.40a) leads to:

$$\begin{aligned} MSE(\bar{y}_{12}) - MSE(\bar{y}_{11}) = \bar{Y}^2 [& \frac{1}{n}(C_{02}+2)(C_{20}-C_{11})^2 \\ & + \frac{1}{n'} C_{20}(C_{20}-C_{11})(C_{20}-C_{11}+2) \\ & + \frac{1}{n^2}\{C_{20}[\frac{\sigma_r^2}{\bar{R}^2} - 3(1-\rho^2)C_{02}] \\ & - 8(C_{20}-C_{11})^2 + 2(C_{12}+C_{30}-2C_{21})\} \\ & + \frac{1}{nn'}\{C_{20}[\frac{\sigma_r^2}{\bar{R}^2}+4C_{11}-3C_{20}-C_{02}]\}] . \end{aligned}$$

(2.40f)

No general conclusions can be drawn from (2.40f). But if approximation (2.28) is valid, then to $O(\frac{1}{n'})$, we deduce that $\bar{y}_{11}$ is superior to $\bar{y}_{12}$ if $\rho < \frac{C_x}{C_y}$.

From equations (2.33b) and (2.35d)

$$MSE(\bar{y}_{13}) - MSE(\bar{y}_{11}) = \bar{Y}^2[\frac{1}{n^2}\{2(2C_{20} + C_{02} - 3C_{11}) + (1 - \frac{\bar{R}}{R})^2 - 2\frac{\bar{R}}{R}(C_{20} - C_{11}) - \frac{2}{\bar{Y}}(\frac{\sigma_{rY}}{R} - \sigma_{rX})\} + \frac{2C_{20}}{nn'}(1 - \frac{\bar{R}}{R})] . \quad (2.41a)$$

As previously mentioned, the difference in MSE's between $\bar{y}_{13}$ and $\bar{y}_{11}$ is of $O(\frac{1}{n^2})$ and in a sufficiently large second-phase sample should be negligible. Under the approximation given by (2.28), we can write expression (2.41a) as:

$$MSE(\bar{y}_{13}) - MSE(\bar{y}_{11}) = \bar{Y}^2[\frac{1}{n^2}\{-(C_{11} - C_{20})^2\} + \frac{2C_{20}}{nn'}(C_{11} - C_{20})] . \quad (2.41b)$$

In this case, $\bar{y}_{13}$ is always superior to $\bar{y}_{11}$ if either

(a) $$\rho < \frac{C_x}{C_y}$$

or

(b) $$\rho > \frac{C_x}{C_y}\left(\frac{2n}{n'C_x^2} + 1\right)$$

We note that under approximation (2.28) therefore $\bar{y}_{13}$ is always superior to $\bar{y}_{11}$ if $\frac{C_x}{C_y} \doteq 1$.

The next pair of estimators we consider are Beale's against the classical ratio estimator.

$$MSE(\bar{y}_{14}) - MSE(\bar{y}_{11}) = \bar{Y}^2\left[\frac{1}{n^2}\{2(C_{12} + C_{30} - 2C_{21}) - 7(C_{20} - C_{11})^2 - 2(1-\rho^2)C_{20}C_{02}\} + \frac{2C_{20}}{nn'}(C_{11} - C_{20})\right]. \quad (2.42a)$$

If $C_{21} = C_{30} = C_{12}$, as in the case of the bivariate normal; then (2.42a) reduces to:

$$MSE(\bar{y}_{14}) - MSE(\bar{y}_{11}) = \bar{Y}^2[\frac{1}{n}\{-7(C_{20}-C_{11})^2 - 2(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{-2C_{20}(C_{20}-C_{11})\}] . \qquad (2.42b)$$

Hence if $\rho \leq \frac{C_x}{C_y}$, then $\bar{y}_{14}$ will always be more efficient than y_{11}. In many repeated surveys in which use is made of ratio-type estimation, one would expect $C_x \doteq C_y$ and hence in this case $\bar{y}_{14}$ will be more precise than $\bar{y}_{11}$. We note that $\rho > \frac{C_x}{C_y}$ does not necessarily imply that $\bar{y}_{11}$ is superior to $\bar{y}_{14}$. We note also that if the term of $O(\frac{1}{nn'})$ is omitted, $\bar{y}_{14}$ is always better than $\bar{y}_{11}$.

From equations (2.33b) and (2.38),

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{11}) = \bar{Y}^2[\frac{1}{n}\{-5(C_{20}-C_{11})^2 - (1-\rho^2)C_{20}C_{02}$$

$$+.2(C_{12}+C_{30}-2C_{21})\}$$

$$+ \frac{1}{nn'}\{2C_{20}(C_{11}-C_{20})\}] . \qquad (2.43)$$

If $C_{12} = C_{30} = C_{21}$ as in the case when (X, Y) follow a bivariate normal distribution, then $\bar{y}_{16}$ is better than $\bar{y}_{11}$ provided that $\rho \leq \frac{C_x}{C_y}$. To $O(\frac{1}{n^2})$ only, $\bar{y}_{16}$ is always better than $\bar{y}_{11}$ provided $C_{21} = C_{30} = C_{12}$.

From equations (2.35d) and (2.38),

$$\begin{aligned} MSE(\bar{y}_{16}) - MSE(\bar{y}_{13}) = \bar{Y}^2[\frac{1}{n^2}\{&-5(C_{20}-C_{11})^2 \\ &- (1-\rho^2)C_{20}C_{02} - 2(2C_{21}-C_{30}-C_{12}) \\ &- 2(2C_{20}-3C_{11}+C_{02}) - (1-\frac{\bar{R}}{R})^2 \\ &+ 2\frac{\bar{R}}{R}(C_{20}-C_{11}) + \frac{2}{\bar{Y}}(\frac{\sigma_{rY}}{R} - \sigma_{rX})\} \\ &+ \frac{2C_{20}}{nn'}\{(C_{11}-C_{20}) - (1-\frac{\bar{R}}{R})\}] . \end{aligned} \qquad (2.44a)$$

Expression (2.44a) does not lend itself to an easy interpretation without further approximations. If we invoke (2.28) and (2.35e), (2.44a) becomes:

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{13}) \doteq \bar{Y}^2[\frac{1}{n}\{-4(C_{20}-C_{11})^2$$

$$-(1-\rho^2)C_{20}C_{02}$$

$$-2(2C_{21}-C_{30}-C_{12})\}] . \qquad (2.44b)$$

The r.h.s. of (2.44b) is clearly negative if $(2C_{21}-C_{30}-C_{12}) \geq 0$. For this case, $\bar{y}_{16}$ will always be superior to $\bar{y}_{13}$.

From equations (2.37) and (2.38),

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{14}) = \bar{Y}^2[\frac{1}{n}\{2(C_{20}-C_{11})^2 + (1-\rho^2)C_{20}C_{02}\}]. \qquad (2.45)$$

The r.h.s. of equation 2.45) is always non-negative and hence $\bar{y}_{14}$ is always to be preferred to $\bar{y}_{16}$. Since $MSE(\bar{y}_{14}) = MSE(\bar{y}_{15})$, an identical argument holds for $\bar{y}_{15}$ vs $\bar{y}_{16}$.

We note that we have not carried out the following comparisons in the general case because they do not lead to any useful results: ($\bar{y}_{13}$ vs $\bar{y}_{12}$), ($\bar{y}_{14}$ vs $\bar{y}_{12}$) and ($\bar{y}_{16}$ vs $\bar{y}_{12}$). Also in the general case, $MSE(\bar{y}_{13}) - MSE(\bar{y}_{14})$ does not yield any fruitful results but when we assume the approximations given by (2.28) and (2.35e), and also take $C_{21} = C_{12} = C_{30}$, we can obtain directly or deduce from (2.44b) and (2.45) that $\bar{y}_{14}$ is always superior to $\bar{y}_{13}$.

To conclude this part of our discussions, we summarize our main results as follows:

(1) We recall that both R and $\bar{R}$ are positive. Hence if we consider terms up to $O(\frac{1}{n'})$ only, the five estimators $\bar{y}_{1j}$ $(j = 1, 3, 4, 5, 6)$ are equally precise and $\bar{y}_{12}$ is inferior to each of them if either

(a) $$\bar{R} \geq R \quad \text{i.e.} \quad \rho_1 \leq 0 .$$

and $$\rho \leq (\frac{\bar{R}+R}{2R})(1 + \frac{n}{n'}) \frac{C_x}{C_y}$$

or (b) $$\bar{R} < R \quad \text{i.e.} \quad \rho_1 > 0$$

and $$\rho \geq (\frac{\bar{R}+R}{2R})(1 + \frac{n}{n'}) \frac{C_x}{C_y} .$$

If $\frac{C_x}{C_y} \doteq 1$ and n' is so large that $(1 + \frac{n}{n'}) \doteq 1$, then condition (a) above reduces simply to

$$\rho_1 \leq 0 \quad \text{and} \quad \rho \leq 1 ,$$

the latter being satisfied in every case.

(2) With respect to the five estimators $\bar{y}_{1j}$ $(j = 1, 3, 4, 5, 6)$

if (a) $$C_{21} = C_{30} = C_{12}$$

(b) the approximations given by (2.28) and (2.35e) are assumed

(c) either $\rho \leq \frac{C_x}{C_y}$ or $\rho > \frac{C_x}{C_y}\left(\frac{2n}{n'C_x^2} + 1\right)$

(The latter condition is only relevant to the comparison $\bar{y}_{13}$ vs $\bar{y}_{11}$)

then, to $O\left(\frac{1}{(n')^2}\right)$,

$$MSE(\bar{y}_{14}) = MSE(\bar{y}_{15}) \leq MSE(\bar{y}_{16}) \leq MSE(\bar{y}_{13}) \leq MSE(\bar{y}_{11}) \qquad (2.46)$$

In such a situation, also, we would prefer $\bar{y}_{14}$ to $\bar{y}_{15}$, since it has a smaller bias.

We note that condition 2(a) above is satisfied when (X, Y) belong to a bivariate normal population. We can obviate the conceptual difficulty arising from our initial assumption that X and Y are both positive by assuming that the parameters of the bivariate normal distribution are such that the probability of either X or Y being non-positive is very small.

We note further that $\bar{y}_{13}$ is superior to $\bar{y}_{11}$ under conditions 2(b) and 2(c) only while $\bar{y}_{14}$ is superior to $\bar{y}_{16}$ in all cases.

We now compare some of the six estimators under assumption 3(b) of section C of this chapter. Under this assumption, the pairs of values (x_i, y_i) satisfy the model:

$$y_i = \alpha + \beta x_i + e_i \qquad (2.47)$$

where $E(e_i | x_i) = 0$ and $E(e_i^2 | x_i) = ax_i^g$, $g \geq 0$, $a > 0$. Thus, under

this model,

$$C_{02} = \frac{\beta^2}{R^2} C_{20} + a \frac{Ex_i^g}{\bar{Y}^2} , \quad \bar{Y} = \alpha + \beta \bar{X}$$

$$R = \frac{\alpha}{\bar{X}} + \beta , \qquad \bar{R} = E(\frac{1}{x}) + \beta ,$$

To simplify our results, we shall take $g = 1$. We have:

$$C_{02} = \frac{\beta^2}{R^2} C_{20} + \frac{a\bar{X}}{\bar{Y}^2} ,$$

$$C_{11} = \frac{\beta}{R} C_{20} , \quad C_{21} = \frac{\beta}{R} C_{30} , \quad C_{12} = \frac{\beta^2}{R^2} C_{30} + \frac{a\bar{X}}{\bar{Y}^2} C_{20}$$

and

$$\sigma_{rX} = \bar{Y} - \bar{R}\bar{X} = [1 - \bar{X} E(\frac{1}{x})] .$$

We note that since $E(\frac{1}{X}) \geq \frac{1}{E(X)}$,

$$\sigma_{rX} \quad \begin{matrix} < 0 \text{ if } \alpha > 0 \\ > 0 \text{ if } \alpha < 0 \end{matrix} .$$

Also, $$\sigma_r^2 = \alpha^2 V(\frac{1}{x}) + a E(\frac{1}{x}) .$$

Now put $V(\frac{1}{x}) = \gamma$ and $E(\frac{1}{x}) = \delta$.

Then, $$\sigma_r^2 = \alpha^2 \gamma + a \delta$$

$$\sigma_{rY} = \alpha \beta(1 - \bar{X} \delta) + a \quad . \tag{2.48}$$

Under model (2.47), we obtain after substituting relevant values from (2.48) in (2.40a),

$$\begin{aligned} MSE(\bar{y}_{12}) - MSE(\bar{y}_{11}) = {} & \frac{1}{n}[\alpha^2(\delta^2 - \frac{1}{\bar{X}^2})]\sigma_X^2 \\ & + \frac{1}{n'}[\alpha^2(\delta^2 - \frac{1}{\bar{X}^2}) + 2\beta\alpha(\delta - \frac{1}{\bar{X}})]\sigma_X^2 \\ & + \frac{1}{n^2}[\alpha^2(2C_{30} - 9C_{20}^2) - a\bar{X}C_{20} \\ & + \alpha^2(1 - \delta\bar{X})^2 + (\alpha^2\delta + a\delta)\sigma_X^2] \end{aligned}$$

$$+ \frac{1}{nn'}[(\alpha^2\gamma + a\delta)\sigma_X^2 - \bar{X}^2(3R-\beta)(R-\beta)C_{20}^2$$

$$- a\bar{X}C_{20}] . \qquad (2.49)$$

It can easily be shown that

(1) If we consider terms up to $O(\frac{1}{n'})$ only, $\bar{y}_{11}$ is superior to $\bar{y}_{12}$ provided either

(a) $\alpha > 0$, $\beta > 0$ (The latter condition is a reasonable assumption since in most surveys in which we are likely to use ratio-type estimation, $\rho > 0$ and this implies $\beta > 0$).

or (b) $\alpha < 0$, $\quad 0 < \beta < \frac{|\alpha|}{\bar{X}}(\frac{n'}{n} + 1)$

or (c) $\alpha < 0$, $\quad \beta > \frac{|\alpha|}{2}(\delta + \frac{1}{\bar{X}})$. (2.50a)

We recall that the results in 1 (a) above are the same for all comparisons with $\bar{y}_{12}$ vs $\bar{y}_{1j}$, j = 1, 3, 4, 5, 6, if only terms to $O(\frac{1}{n'})$ are considered. We note further that condition (2.40b) becomes under model (2.47):

$$\alpha > 0 \quad \text{and} \quad \rho \leq \{\frac{\alpha(\delta + \frac{1}{\bar{X}}) + 2\beta}{2(\frac{\alpha}{\bar{X}} + \beta)}\}(1 + \frac{n}{n'})A ,$$

where $$A = \sqrt{\left(\frac{1}{\frac{\beta^2}{R^2} + \frac{a\bar{X}}{\bar{Y}^2} C_{20}}\right)}$$

It is easy to show that under this model this second condition is superfluous, since for $\alpha > 0$ ρ is necessarily less than $\{\frac{(\delta+1/\bar{X})+2\beta}{2(\ /\bar{X}+\beta)}\}(1+\frac{n}{n'})A$.

(2) If we consider terms up to $O(\frac{1}{n^2})$ only, then

(a) $$\alpha > 0 \ , \quad \beta > 0$$

and (b) $$2C_{30} - 9C_{20}^2 > 0 \qquad (2.50b)$$

are sufficient to ensure the superiority of $\bar{y}_{11}$ over $\bar{y}_{12}$,

(3) If we consider terms to $O(\frac{1}{(n')^2})$, then

(a) $$\alpha > 0 \ , \quad \beta > 0$$

(b) $$2C_{30} - 9C_{20}^2 > 0$$

and (c) $$\{\alpha^2\gamma + a(\delta - \frac{1}{\bar{X}}) - (3R - \beta)(R - \beta)C_{20}\} > 0 \qquad (2.50c)$$

will jointly imply that $\bar{y}_{11}$ is better than $\bar{y}_{12}$.

(4) Finally, in the special case $\alpha = 0$ which implies $R = \beta$, $\bar{y}_{11}$ will always be preferred to $\bar{y}_{12}$.

Again under model (2.47),

$$MSE(\bar{y}_{13}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}[\,^2\{2C_{20}(2 - \bar{X}\delta) + (1 - \bar{X}\delta)(3 - \bar{X}\delta)\}] + \frac{2C_{20}}{nn'}[\alpha(\alpha + \beta\bar{X})(1 - \bar{X}\delta)] . \quad (2.51)$$

We note that:

(1) If $\alpha = 0$, $\bar{y}_{13}$ and $\bar{y}_{11}$ have the same efficiency.

(2) If $\alpha > 0$, $\beta > 0$ and $2 \leq \bar{X}\delta \leq 3$, then $\bar{y}_{13}$ is more efficient than $\bar{y}_{11}$.

(3) If the preliminary sample is so large that the term of $O(\frac{1}{nn'})$ is negligible, then $\bar{y}_{13}$ is to be preferred to $\bar{y}_{11}$ if either

(a) $2 \leq \bar{X}\delta \leq 3$

or (b) $1 \leq \bar{X}\delta < 2$ and $C_{20} < -\frac{(1 - \bar{X}\delta)(3 - \bar{X}\delta)}{2(2 - \bar{X}\delta)}$.

(4) If the approximation given by (2.28) is applicable to this model, implying that $\delta\bar{X} - 1 \doteq C_{20}$, then $\alpha > 0$ would ensure the superiority of $\bar{y}_{13}$ over $\bar{y}_{11}$.

Under model (2.47),

$$MSE(\bar{y}_{14}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}\{\alpha^2(2C_{30} - 7C_{20}^2)\}$$

$$+ \frac{1}{nn'}\{-2\alpha(\alpha + \beta\bar{X})C_{20}^2\} . \qquad (2.52)$$

Hence $\bar{y}_{14}$ is superior to $\bar{y}_{11}$ if $\alpha > 0$ and $2C_{30} < 7C_{20}^2$. No such simple conclusion can be reached if $\alpha < 0$. We note that for the second term to be negative, we need $\alpha > 0$, since $\bar{Y} = \alpha + \beta\bar{X}$ is positive. We note further that if we neglect terms of $O(\frac{1}{nn'})$, the only condition which has to be satisfied to ensure that $\bar{y}_{14}$ is superior to $\bar{y}_{11}$ is $2C_{30} < 7C_{20}^2$.

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}\{\alpha^2(2C_{30} - 5C_{20}^2) - a\bar{X}C_{20}\}$$

$$+ \frac{1}{nn'}\{-2\alpha(\alpha + \beta\bar{X})C_{20}^2\} . \qquad (2.53)$$

Since a is necessarily positive, $\bar{y}_{16}$ is always to be preferred to $\bar{y}_{11}$ under this model, provided $\alpha > 0$ and $2C_{30} < 5C_{20}^2$.

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{13}) = \frac{1}{n}[\alpha^2\{2C_{30} - 5C_{20}^2 - 2C_{20}(2 - \bar{X}\delta)$$

$$- (1 - \bar{X}\delta)(3 - \bar{X}\delta)\} - a\bar{X}C_{20}]$$

$$+ \frac{1}{nn'}[-2aC_{20}(\alpha + \beta\bar{X})(C_{20} + 1 - \bar{X}\delta)] \quad . \qquad (2.54a)$$

Hence if $\alpha > 0$, $\beta > 0$,

$$2C_{30} < \{5C_{20}^2 + 2C_{20}(2 - \bar{X}\delta) + (1 - \bar{X}\delta)(3 - \bar{X}\delta)$$

$$+ a\bar{X}C_{20}\} \qquad (2.54b)$$

and

$$C_{20} + 1 \geq \bar{X}\delta \quad ,$$

then $\bar{y}_{16}$ is superior to $\bar{y}_{13}$. We note that condition (2.54b) is too complicated to be very useful. However, if we assume approximation (2.28) which under this model is $\bar{X}\delta - 1 \doteq C_{20}$, (2.54b) reduces to

$$2C_{30} < 4C_{20}^2 + a\bar{X}C_{20} \, . \qquad (2.54c)$$

$$MSE(\bar{y}_{12}) - MSE(\bar{y}_{14}) = \frac{1}{n}[\alpha^2(\delta^2 - \frac{1}{\bar{X}^2})]\sigma_X^2$$

$$+ \frac{1}{n'}[\alpha^2(\delta^2 - \frac{1}{\bar{X}^2}) + 2\beta\alpha(\delta - \frac{1}{\bar{X}})]\sigma_X^2$$

$$+ \frac{1}{n^2}[\alpha^2\gamma + a\sigma_X^2(\delta - \frac{1}{\bar{X}}) + \alpha^2(1 - \bar{X}\delta)^2$$

$$- 2\alpha^2 C_{20}^2]$$

$$+ \frac{1}{nn'}[\alpha^2\gamma + a\sigma_X^2(\delta - \frac{1}{\bar{X}}) - \alpha^2 C_{20}^2]\,. \qquad (2.55)$$

We recall that to $O(\frac{1}{n'})$, the above difference in MSE's has already been considered in the discussions following (2.49). If we use (2.28), namely $\bar{X}\delta - 1 \doteq C_{20}$, then $\bar{y}_{14}$ is superior to $\bar{y}_{12}$ provided

(i) $\alpha > 0$, $\beta > 0$

(ii) $\gamma > C_{20}^2$, where $\gamma = V(\frac{1}{x})$.

We note that as in the general case, the comparisons ($\bar{y}_{13}$ vs $\bar{y}_{12}$) and ($\bar{y}_{16}$ vs $\bar{y}_{12}$) do not lead to very useful results under this model and will thus not be considered here.

Finally, we recall that (2.45) shows that $MSE(\bar{y}_{16}) - MSE(\bar{y}_{14})$ is always non-negative. It can easily be verified that this is so under model (2.47).

To sum up our results under model (2.47), we recall that one of the common measures of skewness is

$$\gamma_1 = \mu_{30}/\mu_{20}^{\frac{3}{2}} = C_{30}/C_{20}^{\frac{3}{2}} .$$

The inequality $2C_{30} < 5C_{20}^2$ can therefore be written as

$$\gamma_1 < \frac{5}{2} C_x .$$

Hence

(1) If we consider terms up to $O(\frac{1}{n'})$ only, the five estimators $\bar{y}_{1j}$ ($j = 1, 3, 4, 5, 6$), as we have previously found, are equally precise and are always superior to $\bar{y}_{12}$ provided that either

(a) $\alpha > 0 , \quad \beta > 0$

or (b) $\alpha < 0 , \quad 0 < \beta < \frac{|\alpha|}{\bar{X}} (1 + \frac{n'}{n})$

(2) If we consider terms up to $O(\frac{1}{n^2})$ only, then

$$MSE(\bar{y}_{14}) = MSE(\bar{y}_{15}) \leq MSE(\bar{y}_{16}) \leq MSE(\bar{y}_{11})$$

provided that

(a) $\beta > 0$

(b) $\gamma_1 < \frac{5}{2} C_x$

We have omitted $\bar{y}_{13}$ from this ranking because the inequality to be satisfied in the general case is not very satisfactory. In (4) below, we include $\bar{y}_{13}$ in the ranking by invoking approximation (2.28).

(3) If we consider terms up to $O(\frac{1}{(n')^2})$, then we have the same ranking as in (2) above under the same conditions with the addition of > 0.

(4) If we assume the approximation given by (2.28) and also

(a) $\alpha > 0, \quad \beta > 0$

(b) $\gamma_1 < \frac{5}{2} C_x$

(c) $a\bar{X} > C_{20}$

then to $O(\frac{1}{(n')^2})$,

$$MSE(\bar{y}_{14}) = MSE(\bar{y}_{14}) \leq MSE(\bar{y}_{16}) \leq MSE(\bar{y}_{13}) \leq MSE(\bar{y}_{11}).$$

As indicated earlier in assumption 3(c) of Section C of this chapter, our next comparisons will be based on the assumed model $y_i = \alpha + \beta x_i + u_i$, where X has a gamma distribution with parameter m and

$$E(u_i|x_i) = 0 , \quad E(u_i^2|x_i) = n\delta , \text{ where } \delta = O(\tfrac{1}{n}) . \tag{2.56}$$

We have in terms of this model,

$$\overline{Y} = \alpha + \beta m , \quad R = \frac{\alpha}{m} + \beta , \quad \overline{R} = \frac{\alpha}{m-1} + \beta ,$$

$$\overline{R} - R = \frac{\alpha}{m(m-1)} ,$$

$$C_{20} = \frac{1}{m} , \quad C_{30} = \frac{2}{m^2} , \quad C_{02} = \frac{\beta^2 m}{(\alpha + \beta m)^2} + \frac{\delta n}{(\alpha + \beta m)^2} ,$$

$$C_{11} = \frac{\beta}{+\beta m} , \quad C_{21} = \frac{2\beta}{(\alpha+\beta m)m} , \quad C_{12} = \frac{2\beta^2}{(\alpha+\beta m)^2} ,$$

$$\sigma_r^2 = \frac{2}{(m-1)^2(m-2)} + \frac{\delta n}{(m-1)(m-2)} ,$$

$$\sigma_{rX} = -\frac{\alpha}{(m-1)} , \quad \sigma_{rY} = -\frac{\alpha\beta}{m-1} + \frac{\delta n}{m-1} . \tag{2.57}$$

Substituting these values in (2.40a), we have:

$$MSE(\bar{y}_{12}) - MSE(\bar{y}_{11}) = \frac{1}{n}\{\frac{\alpha^2(2m-1)}{m(m-1)^2}\}$$

$$+ \frac{1}{n'}\{\frac{\alpha^2(2m-1)}{m(m-1)^2} + \frac{2\alpha\beta}{m-1}\}$$

$$+ \frac{1}{n^2}\{\frac{\alpha^2(15m-3m^2-10)}{m^2(m-1)(m-2)}$$

$$+ \frac{\delta n(9m-2m^2-6)}{m(m-1)(m-2)}\}$$

$$+ \frac{1}{nn'}\{\frac{\alpha^2(12m^2-2m^3-15m+6)}{m^2(m-1)^2(m-2)} + \frac{(3m-2)\delta n}{m(m-1)(m-2)}\}$$

(2.58)

If $\alpha > 0$, $\beta > 0$, then for n sufficiently large so that terms of $O(\frac{1}{n^2})$ are negligible, $MSE(\bar{y}_{12}) > MSE(\bar{y}_{11})$ for $m > 1$. The same condition could have been obtained by interpreting condition (2.40b) in the light of model (2.56). If $\alpha > 0$, $\beta > 0$, $2 < m < 3.69$, then all terms on the r.h.s. of equation (2.58) are positive, implying that $\bar{y}_{12}$ is inferior to $\bar{y}_{11}$ under these conditions. For $m > 3.69$, the efficiency of $\bar{y}_{11}$ relative to $\bar{y}_{12}$ will depend on the particular values given to α, β, δ, n, and n'.

From (2.41a) and (2.57),

$$MSE(\bar{y}_{13}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}\{\frac{\alpha^2(4 - 3m)}{m(m-1)^2} - \frac{2\delta n}{m-1}\}$$

$$- \frac{1}{nn'}\{\frac{2\alpha(\alpha+\beta m)}{m(m-1)}\} \qquad (2.59)$$

Hence $\alpha > 0$, $\beta > 0$ and $m > \frac{4}{3}$ guarantee that $\bar{y}_{13}$ is better than $\bar{y}_{11}$. If we ignore terms of $O(\frac{1}{nn'})$, then α need not be positive to ensure the superiority of $\bar{y}_{13}$ over $\bar{y}_{11}$, provided, as before, $m > \frac{4}{3}$. We note that approximation (2.28) is not generally applicable here, unless m is very large; for $\sigma_{rX} = \frac{-\alpha}{m-1}$ and $\bar{Y}(C_{11} - C_{20}) = -\frac{\alpha}{m}$. For m sufficiently large so that $\frac{\alpha}{m} \doteq \frac{\alpha}{m-1}$, $\bar{y}_{13}$ is to be preferred to $\bar{y}_{11}$ if $\alpha > 0$.

$$MSE(\bar{y}_{14}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}\{-(\frac{3\alpha^2}{m^2} + \frac{2\delta n}{m})\}$$

$$+ \frac{1}{nn'}\{-\frac{2\alpha(\alpha + \beta m)}{m^2}\}. \qquad (2.60)$$

Hence $\bar{y}_{14}$ is always superior to $\bar{y}_{11}$ under this model if $\alpha > 0$ and $\beta > 0$.

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{11}) = [\frac{1}{n^2}\{\frac{\alpha^2+\delta nm}{m^2}\}$$

$$+ \frac{1}{nn'}\{\frac{2\alpha(\alpha+\beta m)}{m^2}\}] \quad . \qquad (2.61)$$

Again $\bar{y}_{16}$ is better than $\bar{y}_{11}$ for $\alpha > 0$ and $\beta > 0$.

From (2.34f), (2.36b) and (2.57),

$$MSE(\bar{y}_{12}) - MSE(\bar{y}_{14}) = \frac{1}{n}\{\frac{\alpha^2(2m-1)}{m(m-1)^2}\} + \frac{1}{n'}\{\frac{\alpha^2(2m-1)}{m(m-1)^2} + \frac{2\alpha\beta}{m-1}\}$$

$$+ \frac{1}{n^2}\{\frac{2\alpha^2(3m-2)}{m^2(m-1)(m-2)} + \frac{(3m-2)\delta n}{m(m-1)(m-2)}\}$$

$$+ \frac{1}{nn'}\{\frac{\alpha^2(2m^3-4m^2+5m-2)}{m^2(m-1)^2(m-2)} + \frac{(3m-2)\delta n}{m(m-1)(m-2)}\}. \qquad (2.62)$$

We note that $\bar{y}_{14}$ is superior to $\bar{y}_{12}$ if $m > 2$, $\alpha > 0$ and $\beta > 0$. If $\alpha > 0$, $\beta > 0$ but $\frac{1}{2} < m < 2$, the comparison becomes more complicated. In this case, it is possible to choose special values for n , n' and δ for which $\bar{y}_{14}$ will be superior to $\bar{y}_{12}$ and vice versa. For $\alpha > 0$, $\beta > 0$ and $0 < m < \frac{1}{2}$, $\bar{y}_{12}$ is better than $\bar{y}_{14}$.

From (2.34f), (2.38) and (2.57),

$$MSE(\bar{y}_{12}) - MSE(\bar{y}_{16}) = \frac{1}{n}\{\frac{\alpha^2(2m-1)}{m(m-1)^2} + \frac{1}{n'}\{\frac{\alpha^2(2m-1)}{m(m-1)^2} + \frac{2\alpha\beta}{m-1}\}$$

$$+ \frac{1}{nm(m-1)(m-2)}[\frac{(m^2-6m+4)}{n}\{-\frac{2\alpha^2}{m} - \delta n\}$$

$$+ \frac{1}{n'}\{\frac{\alpha^2(4m^2-5m+2)}{m(m-1)} + \delta n(3m-2)\}] \qquad (2.63)$$

The relative efficiency of $\bar{y}_{16}$ with respect to $\bar{y}_{12}$ depends on the value of m. If $2 < m < 5.236$ and $\alpha > 0$, $\beta > 0$, then $\bar{y}_{16}$ is superior to $\bar{y}_{12}$.

As previously stated in the discussion following (2.40a), the difference in MSE's between $\bar{y}_{12}$ and y_{1j} is the same for $j = 1, 3, 4, 5, 6$. This is illustrated by (2.58), (2.59), (2.62) and (2.63).

From (2.35d), (2.36b) and (2.57),

$$MSE(\bar{y}_{13}) - MSE(\bar{y}_{14}) = \frac{1}{n^2}\{\frac{\alpha^2(3-2m)}{m^2(m-1)^2} - \frac{2\delta n}{m(m-1)}\}$$

$$+ \frac{1}{nn'}\{-\frac{2\alpha(\alpha+\beta m)}{m^2(m-1)}\} \qquad (2.64)$$

Since $\bar{Y} = \alpha + \beta m$ is essentially positive, $\bar{y}_{13}$ is superior to $\bar{y}_{14}$ if $\alpha > 0$ and $m > \frac{3}{2}$. If the term of $O(\frac{1}{nn'})$ is negligible, then we only require $m > \frac{3}{2}$ for $\bar{y}_{13}$ to be superior to $\bar{y}_{14}$.

From (2.43) and (2.57),

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{13}) = \frac{1}{n^2}\{\frac{\alpha^2(2m^2 - 2m - 1)}{m^2(m-1)^2} + \frac{\delta n(m+1)}{m(m-1)}\}$$

$$+ \frac{1}{nn'}\{\frac{2\alpha(\alpha+\beta m)}{m^2(m-1)}\} . \qquad (2.65)$$

For $\alpha > 0$, $\beta > 0$ and $m > 1.366$, the r.h.s. of (2.65) is always positive and hence $\bar{y}_{13}$ is to be preferred to $\bar{y}_{16}$ under these conditions.

In concluding our comparisons under model (2.56), we note that:

(1) If n is so large that terms of $O(\frac{1}{n^2})$ are negligible, then the five estimators $\bar{y}_{1j}$ (j = 1, 3, 4, 5, 6) are equally efficient and each is superior to $\bar{y}_{12}$ provided $\alpha > 0$, $\beta > 0$ and $m > 1$.

(2) If we consider terms up to and including those of $O(\frac{1}{nn'})$ and if

(a) $\alpha > 0$, $\beta > 0$

(b) $2 < m < 3.69$

then

$$MSE(\bar{y}_{13}) < MSE(\bar{y}_{14}) = MSE(\bar{y}_{15}) \leq MSE(\bar{y}_{16}) \leq MSE(\bar{y}_{11}) \leq MSE(\bar{y}_{12}) \qquad (2.66)$$

We note all comparisons, except $\bar{y}_{12}$ vs $\bar{y}_{11}$ do not require a condition as strong as 2(b) above. For all $m > 1.5$, the positions of the first five estimators in (2.66) are unchanged relative to each other but for certain special values of δ and m (> 3.69) $\bar{y}_{12}$ could be better than $\bar{y}_{11}$.

The final comparisons to be made under Scheme I relate to the model mentioned in assumption 3(d) of section C of this chapter. We assume that the regression of Y on X is quadratic and that this relationship is given by

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \qquad (2.67)$$

where X is a normal variable with $E(x_i) = 1$ and $V(x_i) = h$, $E(e_i|x_i) = 0$, $V(e_i|x_i) = n\delta$ and $\delta = O(\frac{1}{n})$. We note that if we put $\beta_2 = 0$, we get the linear model used by Durbin [13].

Under model (2.67)

$$\bar{X} = 1 \qquad \bar{Y} = \beta_0 + \beta_1 + \beta_2 + \beta_2 h = \gamma \text{ , say,}$$

$$C_{20} = h \qquad C_{11} = \frac{(\beta_1 + 2\beta_2)h}{\gamma}$$

$$C_{02} = \frac{\beta_1^2 h + 2\beta_2^2 h(h+2) + 4\beta_1\beta_2 h + n\delta}{\gamma^2}$$

$$C_{30} = 0 \qquad C_{21} = \frac{2\beta_2 h^2}{\gamma}$$

$$C_{12} = \frac{4\beta_2 h^2(2\beta_2 + \beta_1)}{\gamma^2} \tag{2.68}$$

Under model (2.67)

$$\begin{aligned} MSE(\bar{y}_{14}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}[&- 7h^2(\gamma - \beta_1 - 2\beta_2)^2 \\ &- 2h(2\beta_2^2 h^2 + n\delta) \\ &- 8\beta_2 h^2(\beta_2 h + \beta_0 - \beta_2)] \\ &+ \frac{1}{nn'}[- h^2\gamma(\beta_2 h + \beta_0 - \beta_2)] \end{aligned} \tag{2.69}$$

Hence if $\beta_i > 0$ $(i = 0, 2)$ and $h > 1 - \frac{\beta_0}{\beta_2}$, then $\bar{y}_{14}$ will be superior to $\bar{y}_{11}$.

From (2.43) and (2.68),

$$MSE(\bar{y}_{16}) - MSE(\bar{y}_{11}) = \frac{1}{n^2}[- 5h^2(\gamma - \beta_1 - 2\beta_2)^2$$

$$- h(2\beta_2^2 h^2 + n\delta)$$

$$- 8\beta_2 h^2(\beta_2 h - \beta_2 + \beta_0)]$$

$$+ \frac{1}{nn'}[- h^2\gamma(\beta_2 h + \beta_0 - \beta_2)] \qquad (2.70)$$

A sufficient condition for $\bar{y}_{16}$ to be superior to $\bar{y}_{11}$ is the same as given above for $\bar{y}_{14}$ to be better than $\bar{y}_{11}$ under the same model (2.67).

Hence under (2.67),

$$MSE(\bar{y}_{14}) = MSE(\bar{y}_{15}) \leq MSE(\bar{y}_{16}) \leq MSE(\bar{y}_{11})$$

if $\beta_i > 0$ $(i = 0, 2)$ and $h > 1 - \frac{\beta_0}{\beta_2}$.

We note that $\bar{y}_{12}$ and $\bar{y}_{13}$ have not been considered in this set of comparisons since $r = \frac{Y}{X}$ (and thus $\bar{R}$) is undefined for $x = 0$, the range of the normal variate being from $-\infty$ to $+\infty$. For $\bar{y}_{11}$, $\bar{y}_{14}$,

$\bar{y}_{15}$ and $\bar{y}_{16}$, we assume that the parameters of the normal distribution are such that $P\{\bar{x}_n \leq 0\}$ is negligible. This is equivalent to assuming that $E(\frac{\bar{y}_n}{\bar{x}_n})$ is well defined.

We also note that if we put $\beta_2 = 0$ in (2.67), the regression of Y on X becomes linear. Under this linear model,

$$MSE(\bar{y}_{14}) = MSE(\bar{y}_{15}) \leq MSE(\bar{y}_{16}) \leq MSE(\bar{y}_{11})$$

if $\beta_0 > 0$. (We recall that $\gamma > 0$ by assumption 1(a) of section C of this chapter.)

G. Scheme II

Under Scheme II, the second phase sample is a simple random subsample of the first phase sample. The assumptions of section C of this chapter still apply. The estimators we shall consider are the dependent second-phase sample equivalents of those considered in section F, with some minor modifications in some coefficients. The six corresponding estimators are:

(1) The classical ratio estimator

$$\bar{y}_{21} = \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_n', \tag{2.71}$$

(2) The Hartley-Ross estimator

$$\bar{y}_{22} = \bar{r}_n \bar{x}_{n'} + \frac{n(n'-1)}{n'(n-1)} (\bar{y}_n - \bar{r}_n \bar{x}_n) \tag{2.72}$$

(3) Pascual's estimator. Two versions of this estimator are considered here. The first is an exact analogue of (2.20), except that $\bar{x}_n$ is now calculated from a subsample of the preliminary sample of size n'. Hence

$$\bar{y}_{23A} = [\frac{n}{n-1} \frac{\bar{y}_n}{\bar{x}_n} - \frac{1}{n-1} \bar{r}_n] \bar{x}_{n'} . \tag{2.73a}$$

The second is a modified form of (2.73a), so that if we make certain assumptions, its bias will be of $O(\frac{1}{n^2})$. The relevant estimator is

$$\bar{y}_{23B} = [\frac{n(n'-1)}{n'(n-1)} \frac{\bar{y}_n}{\bar{x}_n} - \frac{n'-n}{n'(n-1)} \bar{r}_n] \bar{x}_{n'} \tag{2.73b}$$

(4) Beale's estimator

$$\bar{y}_{24} = \frac{\bar{y}_n}{\bar{x}_n} \bar{x}_{n'} [\frac{1 + (\frac{1}{n} - \frac{1}{n'}) \frac{s_{xy}}{\bar{x}_n \bar{y}_n}}{1 + (\frac{1}{n} - \frac{1}{n'}) \frac{s_x^2}{\bar{x}_n^2}}] \tag{2.74}$$

(5) Tin's estimator

$$\bar{y}_{25} = \frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'}\left[1 + \left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{s_{xy}}{\bar{x}_n\bar{y}_n} - \frac{s_x^2}{\bar{x}_n^2}\right)\right] \tag{2.75}$$

and finally

(6) Quenouille's estimator

$$\bar{y}_{26} = 2\hat{R}\bar{x}_{n'} - \frac{1}{2}\left(\hat{R}_1\bar{x}_{n_1'} + \hat{R}_2\bar{x}_{n_2'}\right) \tag{2.76}$$

where we assume that the preliminary sample of size n' is divided at random into two groups of equal size (i.e. $n_1' = n_2' = n'/2$). The estimators of $\bar{X}$ denoted by $\bar{x}_{n'}$, $\bar{x}_{n_1'}$, $\bar{x}_{n_2'}$ are the means from the preliminary sample, computed for the whole and each of the two equal sized groups respectively. A subsample of $n/2$ units is selected from each half of the preliminary sample. $\hat{R}$ is the ordinary ratio estimator calculated from all units of the second-phase sample of size n and $\hat{R}_i$ $(i = 1, 2)$ is the ratio estimator for the ith group calculated from the second phase sample.

We proceed to calculate the biases, variances and MSE's of the estimators listed above. All expressions are, as before, correct to $O\left(\frac{1}{(n')^2}\right)$.

Proceeding as in section F, we can show that

$$E(\bar{y}_{21}) = \bar{Y}[1 + (\frac{1}{n} - \frac{1}{n'})(C_{20} - C_{11})$$

$$+ \frac{1}{n^2}\{(C_{21} - C_{30}) + 3C_{20}(C_{20} - C_{11})\}$$

$$+ \frac{1}{nn'}\{(C_{30} - C_{21}) - 3C_{20}(C_{20} - C_{11})\}] \qquad (2.77)$$

It is easy to show that $\bar{y}_{22}$ is unbiased. The proof can be found in most intermediate sampling textbooks [9].

$$E(\bar{y}_{23A}) = \bar{Y}[1 + (\frac{1}{n} - \frac{1}{n'})(C_{20} - C_{11}) + \frac{1}{n}\frac{\sigma_{rX}}{\bar{Y}}$$

$$+ \frac{1}{n^2}\{(C_{21} - C_{30}) + (C_{20} - C_{11})(3C_{20} + 1) + \frac{\sigma_{rX}}{\bar{Y}}\}$$

$$+ \frac{1}{nn'}\{(C_{30} - C_{21}) - (C_{20} - C_{11})(3C_{20} + 1) - \frac{\sigma_{rX}}{\bar{Y}}\}] \qquad (2.78a)$$

If we assume approximation (2.28) i.e. $\frac{\sigma_{rX}}{\bar{Y}} \doteq (C_{11} - C_{20})$, then (2.78a) becomes:

$$Ea(\bar{y}_{23A}) \doteq \bar{Y}[1 - \frac{1}{n'}(C_{20} - C_{11})$$

$$+ \frac{1}{n^2}\{(C_{21} - C_{30}) + 3C_{20}(C_{20} - C_{11})\}$$

$$+ \frac{1}{nn'}\{(C_{30} - C_{21}) - 3C_{20}(C_{20} - C_{11})\} \qquad (2.78b)$$

Under this condition, $\beta(\bar{y}_{23A})$ if of $O(\frac{1}{n'})$. Hence we consider the modified estimator $\bar{y}_{23B}$.

$$E(\bar{y}_{23B}) = \bar{Y}[1 + (\frac{1}{n} - \frac{1}{n'})(C_{20} - C_{11} + \frac{\sigma_{rX}}{\bar{Y}})$$

$$+ \frac{1}{n^2}\{(C_{21} - C_{30}) + (C_{20} - C_{11})(3C_{20} + 1) + \frac{\sigma_{rX}}{\bar{Y}}\}$$

$$+ \frac{1}{nn'}\{(C_{30} - C_{21}) - (C_{20} - C_{11})(3C_{20} + 2) - \frac{2\sigma_{rX}}{\bar{Y}}\}$$

$$+ \frac{1}{(n')^2}(C_{20} - C_{11} + \frac{\sigma_{rX}}{\bar{Y}})] \qquad (2.78c)$$

If we again assume that $\frac{\sigma_{rX}}{\bar{Y}} \doteq (C_{11} - C_{20})$, then

$$Ea(\bar{y}_{23B}) \doteq \bar{Y}\left[1 + \frac{1}{n^2}\{(C_{21} - C_{30}) + 3C_{20}(C_{20} - C_{11})\}\right.$$

$$\left. + \frac{1}{nn'}\{(C_{30} - C_{21}) - 3C_{20}(C_{20} - C_{11})\}\right] \qquad (2.78d)$$

$$E(\bar{y}_{24}) = \bar{Y}\left[1 + \frac{1}{n^2}\{2(C_{30} - C_{21}) - 2C_{20}(C_{20} - C_{11})\}\right]$$

$$+ \frac{1}{nn'}\{3(C_{21} - C_{30}) + 4C_{20}(C_{20} - C_{11})\}$$

$$\left. + \frac{1}{(n')^2}\{(C_{30} - C_{21}) - 2C_{20}(C_{20} - C_{11})\}\right] \qquad (2.79)$$

$$E(\bar{y}_{25}) = \bar{Y}\left[1 + \frac{1}{n^2}\{2(C_{30} - C_{21}) - 3C_{20}(C_{20} - C_{11})\}\right.$$

$$+ \frac{1}{nn'}\{3(C_{21} - C_{30}) + 6C_{20}(C_{20} - C_{11})\}$$

$$\left. + \frac{1}{(n')^2}\{(C_{30} - C_{21}) - 3C_{20}(C_{20} - C_{11})\}\right] \qquad (2.80)$$

Finally

$$E(\bar{y}_{26}) = \bar{Y}[1 - \frac{2}{n^2}\{(C_{21} - C_{30}) + 3C_{20}(C_{20} - C_{11})\}$$

$$+ \frac{2}{nn'}\{(C_{21} - C_{30}) + 3C_{20}(C_{20} - C_{11})\}] \qquad (2.81)$$

From these results, $\bar{y}_{22}$ is the only unbiased estimator. However, we can rank some of the estimators if we make the following assumptions:

1. If n is sufficiently large so that terms of $O(\frac{1}{n^2})$ are negligible, then $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$ have negligible biases and

$$\beta(\bar{y}_{21}) < \beta(\bar{y}_{23B}) < \beta(\bar{y}_{23A})$$

if

(a) $\rho < \frac{C_x}{C_y}$ and $\frac{\sigma_{rX}}{\bar{Y}} > 0$

or (b) $\rho > \frac{C_x}{C_y}$ and $\frac{\sigma_{rX}}{\bar{Y}} < 0$.

Using approximation for $\frac{\sigma_{rX}}{\bar{Y}}$ (2.28), the two pairs of inequalities in (a) and (b) appear contradictory (i.e. not likely to occur). If we reverse

one of the inequalities in both (a) and (b) above, the comparison of the bias becomes less straightforward. For example,

$$\beta(\bar{y}_{23A}) < \beta(\bar{y}_{21})$$

if either

(i) $$\rho < \frac{C_x}{C_y}, \quad \sigma_{rX} < 0 \quad \text{and} \quad \left|\frac{\sigma_{rX}}{\bar{Y}}\right| < 2\left(1 - \frac{n}{n'}\right)(C_{20} - C_{11})$$

or (ii) $$\rho > \frac{C_x}{C_y}, \quad \sigma_{rX} > 0 \quad \text{and} \quad \frac{\sigma_{rX}}{\bar{Y}} < 2\left(1 - \frac{n}{n'}\right)\left|C_{20} - C_{11}\right|$$

Similarly,

$$\beta(\bar{y}_{23B}) < \beta(\bar{y}_{21})$$

if either

(i) $$\rho < \frac{C_x}{C_y}, \quad \sigma_{rX} < 0 \quad \text{and} \quad \left|\frac{\sigma_{rX}}{\bar{Y}}\right| < 2(C_{20} - C_{11})$$

or (ii) $$\rho > \frac{C_x}{C_y}, \quad \sigma_{rX} > 0 \quad \text{and} \quad \frac{\sigma_{rX}}{\bar{Y}} < 2\left|C_{20} - C_{11}\right|$$

We note that the results for the comparison of the biases of $\bar{y}_{21}$ and $\bar{y}_{23B}$ are identical to those obtained for the corresponding estimators under Scheme I.

2. Again if we omit terms of $O(\frac{1}{n^2})$ and assume approximation (2.28), then $\beta_a(\bar{y}_{23B}) = 0$ and hence

$$|\beta_a(\bar{y}_{23B})| < |\beta_a(\bar{y}_{23A})| < |\beta(\bar{y}_{21})|$$

if $n < \frac{n'}{2}$ and $\rho \neq \frac{C_x}{C_y}$. If $\rho = \frac{C_x}{C_y}$, the biases of the three estimaters are all equal under the above assumptions.

3. We now consider the relative biases of the three estimators whose biases are of $O(\frac{1}{n^2})$: $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$. It is easy to verify that if $C_{21} = C_{30}$ (as in a bivariate normal population), then, to $O(\frac{1}{n^2})$,

$$|\beta(\bar{y}_{24})| \leq |\beta(\bar{y}_{25})| \leq |\beta(\bar{y}_{26})|$$

with equality only if $C_{20} = C_{11}$. These results are similar to those obtained in Section F of this chapter.

4. If approximation (2.28) is applicable, then we can rank $\bar{y}_{23B}$, $\bar{y}_{24}$, $\bar{y}_{25}$, $\bar{y}_{26}$ as follows with respect to bias to $O(\frac{1}{n^2})$

$$|\beta(\bar{y}_{24})| \leq |\beta_a(\bar{y}_{23B})| = |\beta(\bar{y}_{25})| \leq |\beta(\bar{y}_{26})| ,$$

again with the biases equal only if $C_{20} = C_{11}$.

We next consider the variances of the six estimators. It is easy to show that

$$\begin{aligned} V(\bar{y}_{21}) = \bar{Y}^2 [& \frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20}) \\ & + \frac{1}{n^2}(4C_{21} - 2C_{12} + 5C_{11}^2 + 3C_{20}C_{02} \\ & - 2C_{30} - 16C_{20}C_{11} + 8C_{20}^2) \\ & + \frac{1}{nn'}(4C_{30} - 6C_{21} + 22C_{20}C_{11} + 2C_{12} \\ & - 6C_{11}^2 - 3C_{20}C_{02} - 13C_{20}^2) \\ & + \frac{1}{(n')^2}(2C_{21} + C_{11}^2 - 2C_{30} - 6C_{20}C_{11} \\ & + 5C_{20}^2)] \end{aligned} \qquad (2.82a)$$

Hence from (2.77) and (2.82a), we obtain:

$$\begin{aligned} MSE(\bar{y}_{21}) = \bar{Y}^2 [& \frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20}) \\ & + \frac{1}{n^2}(4C_{21} - 2C_{12} + 6C_{11}^2 + 3C_{20}C_{02} \\ & - 2C_{30} - 18C_{20}C_{11} + 9C_{20}^2) \\ & + \frac{1}{nn'}(4C_{30} - 6C_{21} + 26C_{20}C_{11} + 2C_{12} - 8C_{11}^2 \\ & - 3C_{20}C_{02} - 15C_{20}^2) \\ & + \frac{1}{(n')^2}(2C_{21} + 2C_{11}^2 - 2C_{30} - 8C_{21}C_{11} + 6C_{20}^2)] . \end{aligned} \tag{2.82b}$$

To obtain $V(\bar{y}_{22})$, we write:

$$\bar{y}_{22} = \bar{r}_n \bar{x}_{n'} + \frac{n'-1}{n'} \bar{Z}_n , \quad \text{where } \bar{Z}_n = \frac{n}{n-1}(\bar{y}_n - \bar{r}_n \bar{x}_n) .$$

Hence

$$V(\bar{y}_{22}) = V(\bar{r}_n \bar{x}_{n'}) + \frac{(n'-1)^2}{(n')^2} V(\bar{Z}_n)$$

$$+ \frac{2n(n'-1)}{n'(n-1)} \mathrm{Cov}(\bar{r}_n \bar{x}_{n'}, \bar{y}_n - \bar{r}_n \bar{x}_n) . \qquad (2.83a)$$

After some elementary but tedious computations, (2.83a) reduces to:

$$V(\bar{y}_{22}) = \frac{1}{n}(\sigma_Y^2 + \bar{R}^2\sigma_X^2 - 2\bar{R}\sigma_{XY}) + \frac{1}{n'}(2\bar{R}\sigma_{XY} - \bar{R}^2\sigma_X^2)$$

$$+ \frac{1}{n^2}(\sigma_r^2\sigma_X^2 + \sigma_{rX}^2 - 2\bar{X}\,E[(\Delta^2 r)\Delta x] + 2\bar{X}\sigma_{rY}$$

$$- 2\bar{X}^2\sigma_r^2 - 2\bar{R}\bar{X}\sigma_{rX})$$

$$+ \frac{1}{nn'}(2\bar{X}^2\sigma_r^2 - 2E[(\Delta^2 x)\Delta^2 r)]$$

$$- 4\bar{R}E(\Delta r)\Delta^2 x - \sigma_r^2\sigma_X^2$$

$$+ 2\bar{R}\sigma_{XY} + 2E(\Delta x \Delta y \Delta r)$$

$$- 2\bar{R}^2\sigma_X^2 - 2\bar{X}\sigma_{rY})$$

$$+ \frac{1}{(n')^2}(2\bar{R}E[(\Delta^2 x)(\Delta r) + \sigma_{rX}^2 - 2\bar{R}\sigma_{XY}$$

$$+ 2\bar{R}^2\sigma_X^2 + 2\bar{R}\bar{X}\sigma_{rX}) \quad . \tag{2.83b}$$

After some straight-forward algebra, (2.83b) reduces further to:

$$V(\bar{y}_{22}) = \frac{1}{n}(\sigma_Y^2 + \bar{R}^2\sigma_X^2 - 2\bar{R}\sigma_{XY}) + \frac{1}{n'}(2\bar{R}\sigma_{XY} - \bar{R}^2\sigma_X^2)$$

$$+ \frac{1}{n^2}(\sigma_r^2\sigma_X^2 + \sigma_{rX}^2) + \frac{1}{nn'}(-2\sigma_{rX}^2 - \sigma_r^2\sigma_X^2)$$

$$+ \frac{1}{(n')^2}(\sigma_{rX}^2) \quad . \tag{2.83c}$$

We note that in Sukhatme's work [54], where the variance is calculated to $O(\frac{1}{n'})$, the second term on the r.h.s. of (2.83c) is written as

$$\frac{1}{n'}\{\bar{R}^2\sigma_X^2 + 2\bar{R}\bar{X}\sigma_{rX} + 2\bar{R}E(\Delta^2 X\Delta r)\} \quad .$$

By writing $E(\Delta^2 X\Delta r)$ as $\sigma_{XY} - \bar{X}\bar{Y} + \bar{X}^2\bar{R} - \bar{R}\sigma_X^2$, it can be verified that the two expressions are identical.

Using approximation (2.28), we obtain:

$$V_a(\bar{y}_{22}) \doteq \bar{Y}^2\left[\frac{1}{n}\{(C_{02} - 2C_{11} + C_{20}) + (C_{20} - C_{11})^2(2 + C_{20})\}\right.$$

$$+ \frac{1}{n'}\{(2C_{11} - C_{20}) - (C_{20} - C_{11})^2(2 + C_{20})\}$$

$$+ \frac{1}{n^2}\{(C_{20} - C_{11})^2 + \frac{\sigma_r^2}{R^2} C_{20}\}$$

$$+ \frac{1}{nn'}\{-2(C_{20} - C_{11})^2 - \frac{\sigma_r^2}{R^2} C_{20}\}$$

$$\left. + \frac{1}{(n')^2}\{(C_{20} - C_{11})^2\}\right] . \qquad (2.83d)$$

Since $\bar{y}_{22}$ is an unbiased estimator, we have:

$$MSE(\bar{y}_{22}) = V(\bar{y}_{22}) \qquad (2.83e)$$

Next, we note that

$$V(\bar{y}_{23A}) = \frac{1}{(n-1)^2}[n^2 V(\frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'}) + V(\bar{r}_n\bar{x}_{n'})$$

$$- 2n\,\mathrm{Cov}(\frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'},\ \bar{r}_n\bar{x}_{n'})] \qquad (2.84a)$$

Since our results are being given to $O(\frac{1}{(n')^2})$ and $V(\bar{r}_n\bar{x}_{n'})$ is of $O(\frac{1}{n})$, it need not be considered. Again, we need to determine $\mathrm{Cov}(\frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'},\ \bar{r}_n\bar{x}_{n'})$ only to $O(\frac{1}{n})$.

It is easy to show that to $O(\frac{1}{n})$,

$$\mathrm{Cov}(\frac{\bar{y}_n}{\bar{x}_n}\bar{x}_{n'},\ \bar{r}_n\bar{x}_{n'}) = \frac{1}{n}(\bar{X}\sigma_{rY} - \bar{Y}\sigma_{rX}) + \frac{1}{n'}(\bar{R}\sigma_{XY} + \bar{Y}\sigma_{rX}) \qquad (2.84b)$$

Hence from (2.82a), (2.84a) and (2.84b),

$$V(\bar{y}_{23A}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})$$

$$+ \frac{1}{n^2}\{2(2C_{21} - C_{12} - C_{30}) + 8(C_{20} - C_{11})^2$$

$$+ 3(1-\rho^2)C_{20}C_{02} + 2(C_{02} - 2C_{11} + C_{20})$$

$$+ \frac{2}{\bar{Y}}(\sigma_{rX} - \frac{\sigma_{rY}}{R})\}$$

$$+ \frac{1}{nn'}\{2(2C_{30} - 3C_{21} + C_{12}) - 13(C_{20} - C_{11})^2$$

$$- 4C_{11}(C_{20} - C_{11}) - 3(1-\rho^2)C_{20}C_{02}$$

$$+ 2(2C_{11} - C_{20}) - \frac{2}{\bar{Y}}(\sigma_{rX} + \frac{\bar{R}\sigma_{XY}}{\bar{Y}})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21} - C_{30}) + 5(C_{20} - C_{11})^2$$

$$+ 4C_{11}(C_{20} - C_{11})\}] \quad . \qquad (2.84c)$$

Hence from (2.78a) and (2.84c) ,

$$MSE(\bar{y}_{23A}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})$$

$$+ \frac{1}{n^2}\{2(2C_{21} - C_{12} - C_{30}) + 9(C_{20} - C_{11})^2$$

$$+ 3(1-\rho^2)C_{20}C_{02} + 2(C_{02} - 2C_{11} + C_{20})$$

$$+ \frac{2\sigma_{rX}}{\bar{Y}}(1 + C_{20} - C_{11}) + \frac{1}{\bar{Y}^2}(\sigma_{rX}^2 - 2\bar{X}\sigma_{rY})\}$$

$$+ \frac{1}{nn'}\{2(2C_{30} - 3C_{21} + C_{12}) - 15(C_{20} - C_{11})^2$$

$$- 3(1-\rho^2)C_{20}C_{02} - 4C_{11}(C_{20} - C_{11})$$

$$- \frac{2\sigma_{rX}}{\bar{Y}}(1 + C_{20} - C_{11}) - \frac{2\bar{R}\sigma_{XY}}{\bar{Y}^2}$$

$$+ 2(2C_{11} - C_{20})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21} - C_{30}) + 6(C_{20} - C_{11})^2$$

$$+ 4C_{11}(C_{20} - C_{11})\}] \quad . \tag{2.84d}$$

If we now use approximations (2.28) and (2.35e), we obtain

$$MSE_a(\bar{y}_{23A}) \doteq \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})$$

$$+ \frac{1}{n^2}\{2(2C_{21} - C_{12} - C_{30}) + 8(C_{20} - C_{11})^2$$

$$+ 3(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{2(2C_{30} - 3C_{21} + C_{12}) - 13(C_{20} - C_{11})^2$$

$$- 3(1-\rho^2)C_{20}C_{02} - 6C_{11}(C_{20} - C_{11})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21} - C_{30}) + 6(C_{20} - C_{11})^2$$

$$+ 4C_{11}(C_{20} - C_{11})\}] . \qquad (2.84e)$$

We next find the variance of the modified Pascual estimator $\bar{y}_{23B}$, using the same approach as for $\bar{y}_{23A}$.

$$V(\bar{y}_{23B}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})$$

$$+ \frac{1}{n^2}\{2(C_{02} - 2C_{11} + C_{20}) + 2(2C_{21} - C_{30} - C_{12})$$

$$+ 8(C_{20} - C_{11})^2 + 3(1-\rho^2)C_{20}C_{02}$$

$$+ \frac{2}{\bar{Y}}(\sigma_{rX} - \frac{\sigma_{rY}}{R})\}$$

$$+ \frac{1}{nn'}\{2(2C_{30} - 3C_{21} + C_{12}) - 13(C_{20} - C_{11})^2$$

$$- 3(1-\rho^2)\, C_{20}C_{02} - 4C_{11}(C_{20} - C_{11})$$

$$+ 2(4C_{11} - 2C_{20} - C_{02})$$

$$+ \frac{2}{\bar{Y}}\left(\frac{\sigma_{rY}}{R} - 2\sigma_{rX} - \bar{R}\bar{X}C_{11}\right)\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21} - C_{30}) + 5(C_{20} - C_{11})^2$$

$$+ 4C_{11}(C_{20} - C_{11}) + 2(C_{20} - 2C_{11})$$

$$+ \frac{2}{\bar{Y}}(\sigma_{rX} + \bar{R}\bar{X}C_{11})\}\Big] . \qquad (2.84f)$$

Using approximations (2.28) and (2.35e), we obtain:

$$V_a(\bar{y}_{23B}) \doteq \bar{Y}^2\left[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})\right.$$

$$+ \frac{1}{n^2}\{2(2C_{21} - C_{12} - C_{30}) + 8(C_{20} - C_{11})^2$$

$$+ 3(1-\rho^2)\, C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{2(2C_{30} - 3C_{21} + C_{12}) - 13(C_{20} - C_{11})^2$$

$$- 3(1-\rho^2)C_{20}C_{02} - 6C_{11}(C_{20} - C_{11})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21} - C_{30}) + 5(C_{20} - C_{11})^2$$

$$+ 6C_{11}(C_{20} - C_{11})\}] . \qquad (2.84g)$$

Under these special conditions,

$$MSE_a(\bar{y}_{23B}) = V_a(\bar{y}_{23B}) \qquad (2.84h)$$

In the general case,

$$MSE(\bar{y}_{23B}) = \bar{Y}^2[\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})$$

$$+ \frac{1}{n^2}\{2(C_{02} - 2C_{11} + C_{20}) + 2(2C_{21} - C_{12} - C_{30})$$

$$+ 9(C_{20} - C_{11})^2 + 3(1-\rho^2)C_{20}C_{02}$$

$$+ \frac{1}{\bar{Y}^2}(\sigma_{rX}^2 - 2\bar{X}\sigma_{rY}) + \frac{2\sigma_{rX}}{\bar{Y}}(1 + C_{20} - C_{11})\}$$

$$+ \frac{1}{nn'}\{2(2C_{30} - 3C_{21} + C_{12}) - 15(C_{20} - C_{11})^2$$

$$- 3(1-\rho^2)\, C_{20}C_{02} - 4C_{11}(C_{20} - C_{11})$$

$$+ 2(4C_{11} - 2C_{20} - C_{02}) - \frac{4\sigma_{rX}}{\bar{Y}}(1 + C_{20} - C_{11})$$

$$- \frac{2}{\bar{Y}^2}(\sigma_{rX}^2 + \bar{R}\sigma_{XY} - \bar{X}\sigma_{rY})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21} - C_{30}) + 6(C_{20} - C_{11})^2$$

$$+ 4C_{11}(C_{20} - C_{11})$$

$$+ 2(C_{20} - 2C_{11}) + \frac{2}{\bar{Y}}\sigma_{rX}(1 + C_{20} - C_{11})$$

$$+ \frac{1}{\bar{Y}^2}(2\bar{R}\sigma_{XY} + \sigma_{rX}^2)\}] \qquad (2.84i)$$

From (2.84d) and (2.84i),

$$\mathrm{MSE}(\bar{y}_{23A})-\mathrm{MSE}(\bar{y}_{23B}) = \bar{Y}^2\Big[\frac{1}{nn'}\{2(C_{02}-2C_{11}+C_{20}) + \frac{2\sigma_{rX}}{\bar{Y}}(1+C_{20}-C_{11}) - \frac{2}{\bar{Y}^2}(\bar{X}\sigma_{rY}-\sigma_{rX}^2)\} + \frac{1}{(n')^2}\{2(2C_{11}-C_{20}) - \frac{2\sigma_{rX}}{\bar{Y}}(1+C_{20}-C_{11}) - \frac{1}{\bar{Y}^2}(2\bar{R}\sigma_{XY}+\sigma_{rX}^2)\}\Big] \qquad (2.84j)$$

Under approximations (2.28) and (2.35e), this reduces to

$$\mathrm{MSE}_a(\bar{y}_{23A}) - \mathrm{MSE}_a(\bar{y}_{23B}) \doteq \bar{Y}^2\Big[\frac{1}{(n')^2}(C_{20}-3C_{11})(C_{20}-C_{11})\Big] \qquad (2.84k)$$

We recall from section C of this chapter that $\rho > 0$ and $C_{11} > 0$. Hence

$$(C_{20}-3C_{11})(C_{20}-C_{11}) < 0 \quad \text{if} \quad \frac{1}{3}\frac{C_x}{C_y} < \rho < \frac{C_x}{C_y} \; .$$

For large n' which is usual in double sampling schemes, this difference (2.84k) is very small. Hence, in our subsequent comparisons, we shall use $\bar{y}_{23B}$ only.

To obtain the variance of $\bar{y}_{24}$, we write $s_x^2 = 1 + \delta s_x^2$, $\bar{x}_n = 1 + \delta \bar{x}_n$ etc. After some simplification, we obtain:

$$V(\bar{y}_{24}) = \bar{Y}^2 [\frac{1}{n}(C_{02} - 2C_{11} + C_{20}) + \frac{1}{n'}(2C_{11} - C_{20})$$

$$+ \frac{1}{n^2}\{2(C_{20} - C_{11})^2$$

$$+ (1-\rho^2)\, C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{-(C_{20} - C_{11})^2$$

$$- (1-\rho^2)\, C_{20}C_{02}\}$$

$$+ \frac{1}{(n')^2}\{-(C_{20} - C_{11})^2\}] \qquad (2.85a)$$

Since $\beta(\bar{y}_{24})$ is of $O(\frac{1}{n^2})$, we have:

$$MSE(\bar{y}_{24}) = V(\bar{y}_{24}) \tag{2.85b}$$

Similarly it can be shown that

$$\begin{aligned} MSE(\bar{y}_{25})=MSE(\bar{y}_{24}) = \bar{Y}^2[&\tfrac{1}{n}(C_{02}-2C_{11}+C_{20}) \\ &+\tfrac{1}{n'}(2C_{11}-C_{20}) \\ &+\tfrac{1}{n^2}\{2(C_{20}-C_{11})^2+(1-\rho^2)C_{20}C_{02}\} \\ &+\tfrac{1}{nn'}\{-(C_{20}-C_{11})^2-(1-\rho^2)C_{20}C_{02}\} \\ &+\tfrac{1}{(n')^2}\{-(C_{20}-C_{11})^2\}] \end{aligned} \tag{2.86}$$

$$\begin{aligned} V(\bar{y}_{26}) = \bar{Y}^2[&\tfrac{1}{n}(C_{02}-2C_{11}+C_{20})+\tfrac{1}{n'}(2C_{11}-C_{20}) \\ &+\tfrac{1}{n^2}\{4(C_{20}-C_{11})^2+2(1-\rho^2)C_{20}C_{02}\} \end{aligned}$$

$$+ \frac{1}{nn'}\{-10(C_{20}-C_{11})^2 - 2(1-\rho^2)\, C_{20}C_{02}\}$$

$$+ \frac{1}{(n')^2}\{6(C_{20}-C_{11})^2\}] \qquad (2.87a)$$

Since $\beta(\bar{y}_{26})$ is of $O(\frac{1}{n^2})$,

$$MSE(\bar{y}_{26}) = V(\bar{y}_{26}) \qquad (2.87b)$$

We note that here, as in Scheme I, with the exception of $\bar{y}_{22}$, all the remaining estimators have the same MSE's to $O(\frac{1}{n'})$. Hence for proper comparisons we need to take into account terms of $O(\frac{1}{(n')^2})$.

As in Section F, we shall compare the MSE's of the six estimators in this section under four different assumptions. First, we make no assumptions about the distribution of the (x_i, y_i). Thus, from (2.82b) and (2.83e), we have

$$MSE(\bar{y}_{22}) - MSE(\bar{y}_{21}) = \bar{Y}^2[\frac{1}{n}\{\frac{\bar{R}^2}{R^2} - 1)\, C_{20} - 2(\frac{\bar{R}}{R} - 1)\, C_{11}\}$$

$$+ \frac{1}{n'}\{2(\frac{\bar{R}}{R} - 1)\, C_{11} - (\frac{\bar{R}^2}{R^2} - 1)C_{20}\}$$

$$+ \frac{1}{n^2}\{\frac{\sigma_r^2}{R^2} C_{20} + \frac{\sigma_{rX}^2}{\bar{Y}^2} + 2(C_{30} + C_{12} - 2C_{21})$$

$$- 9(C_{20} - C_{11})^2 - 3(1-\rho^2) C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{- \frac{\sigma_r^2}{R^2} C_{20} - \frac{2\sigma_{rX}^2}{\bar{Y}^2} + 2(3C_{21} - 2C_{30} - C_{12})$$

$$+ (5C_{20} - 2C_{11})(3C_{20} - 4C_{11}) + 3C_{20}C_{02}\}$$

$$+ \frac{1}{(n')^2}\{\frac{\sigma_{rX}^2}{\bar{Y}^2} - 2(C_{12} - C_{30})$$

$$- 2(3C_{20} - C_{11})(C_{20} - C_{11})\}] . \qquad (2.88a)$$

Using approximation (2.28), we have:

$$MSE_a(\bar{y}_{22}) - MSE(\bar{y}_{21}) = \bar{Y}^2[(\frac{1}{n} - \frac{1}{n'})(C_{20} - C_{11})^2(2 + C_{20})$$

$$+ \frac{1}{n^2}\{2(C_{30} + C_{12} - 2C_{21}) - 8(C_{20} - C_{11})^2$$

$$- 3(1-\rho^2) C_{20}C_{02} + \frac{\sigma_r^2}{R^2} C_{20}\}$$

$$+ \frac{1}{nn'}\{2(3C_{21} - C_{12} - 2C_{30}) + 13(C_{20} - C_{11})^2$$

$$+ 3(1-\rho^2) C_{20}C_{02} + 4C_{11}(C_{20} - C_{11}) - \frac{\sigma_r^2}{R^2} C_{20}\}$$

$$+ \frac{1}{(n')^2}\{2(C_{30} - C_{21}) - 5(C_{20} - C_{11})^2$$

$$- 4C_{11}(C_{20} - C_{11})\}] \quad . \qquad (2.88b)$$

We deduce from Section C that $\bar{X}$, $\bar{Y}$, R and $\bar{R}$ are all positive as previously stated. We also recall from Chapter I that Sukhatme has compared the MSE's of $\bar{y}_{22}$ and $\bar{y}_{21}$ to $O(\frac{1}{n'})$. From (2.88a), we deduce that to $O(\frac{1}{n'})$,

$$MSE(\bar{y}_{22}) < MSE(\bar{y}_{21})$$

if either

$$\bar{R} > R \quad \text{and} \quad \rho > \frac{(\bar{R} + R)}{2} \frac{\sigma_X}{\sigma_Y}$$

or $$\bar{R} < R \quad \text{and} \quad \rho < \frac{(\bar{R} + R)}{2} \frac{\sigma_X}{\sigma_Y} \qquad (2.88c)$$

We note that to this order of approximation, $\bar{y}_{22}$ and $\bar{y}_{21}$ are equally efficient if $\bar{R} = R$. As previously stated if x_i is the value of y_i at some previous time, then $\frac{C_x}{C_y} \doteq 1$ is likely to be a reasonable assumption. Hence, under this assumption, the first pair of inequalities in (2.83c) become $\bar{R} > R$ and $\rho > 1$, the second inequality being clearly impossible. Hence only the second pair of inequalities will be valid in this case.

We observe that to $O(\frac{1}{n'})$, the same conclusions hold when $\bar{y}_{22}$ is compared to $\bar{y}_{2j}$, $j = 3, 4, 5, 6$.

From (2.82b) and (2.84i),

$$MSE(\bar{y}_{23B}) - MSE(\bar{y}_{21}) = \bar{Y}^2 [\frac{1}{n^2} \{2(C_{02} - 2C_{11} + C_{20})$$

$$+ \frac{2\sigma_{rX}}{\bar{Y}} (1 + C_{20} - C_{11}) + \frac{1}{\bar{Y}^2} (\sigma^2_{rX} - 2\bar{X}\sigma_{rY})\}$$

$$+\frac{1}{nn'}\{2(4C_{11}-2C_{20}-C_{02})$$

$$+\frac{2}{\bar{Y}^2}(\bar{X}\sigma_{rY}-\sigma_{rX}^2-\bar{R}\sigma_{XY})$$

$$-\frac{4\sigma_{rX}}{\bar{Y}}(1+C_{20}-C_{11})\}$$

$$+\frac{1}{(n')^2}\{2(C_{20}-2C_{11})+\frac{1}{\bar{Y}^2}(2\bar{R}\sigma_{XY}+\sigma_{rX}^2)$$

$$+\frac{2\sigma_{rX}}{\bar{Y}}(1+C_{20}-C_{11})\}]\,. \qquad (2.89a)$$

Under approximations (2.28) and (2.35e), (2.89a) reduces to:

$$MSE_a(\bar{y}_{23B})-MSE(\bar{y}_{21}) = -\bar{Y}^2(\frac{1}{n}-\frac{1}{n'})[\{(\frac{1}{n}-\frac{1}{n'})C_{20}$$

$$-(\frac{1}{n}-\frac{3}{n'})C_{11}\}(C_{20}-C_{11})]\,. \qquad (2.89b)$$

Hence

$$MSE_a(\bar{y}_{23B}) < MSE(\bar{y}_{21}) \quad \text{if} \quad \rho < \frac{C_x}{C_y} \ .$$

Also

$$MSE(\bar{y}_{21}) < MSE_a(\bar{y}_{23B})$$

if

$$\frac{C_x}{C_y} < \rho < \frac{(\frac{1}{n} - \frac{1}{n'})}{(\frac{1}{n} - \frac{3}{n'})} \frac{C_x}{C_y} \ .$$

If we consider terms of $O(\frac{1}{n^2})$ only, then $MSE_a(\bar{y}_{23B}) < MSE(\bar{y}_{21})$ always.

We recall from (2.86) that $MSE(\bar{y}_{24}) = MSE(\bar{y}_{25})$. An alternative approach to the derivation of the variances of $\bar{y}_{24}$ and $\bar{y}_{25}$ establishes this identity more directly. If $Ps_{n'}$ denotes a particular preliminary sample of size n', then

$$V(\bar{y}_{24}) = V[E(\bar{y}_{24}|Ps_{n'})] + E[V(\bar{y}_{24}|Ps_{n'})] \ . \qquad (2.90a)$$

Similarly,

$$V(\bar{y}_{25}) = V[E(\bar{y}_{25}|Ps_{n'})] + E[V(\bar{y}_{25}|Ps_{n'})] \ . \qquad (2.90b)$$

Tin [58] has, shown that

$$V(\bar{y}_{24}|Ps_{n'}) = V(\bar{y}_{25}|Ps_{n'}) \ . \tag{2.91a}$$

To $O(\frac{1}{(n')^2})$,

$$V[E(\bar{y}_{24}|Ps_{n'})] = V[E(\bar{y}_{25}|Ps_{n'})] \tag{2.91b}$$

Hence from (2.90a), ..., (2.91b),

$$V(\bar{y}_{24}) = V(\bar{y}_{25}) \tag{2.92}$$

and

$$MSE(\bar{y}_{24}) = MSE(\bar{y}_{25}) \ .$$

$$MSE(\bar{y}_{25}) - MSE(\bar{y}_{21}) = \bar{Y}^2[\frac{1}{n^2}\{2(C_{30}+C_{12}-2C_{21})$$

$$- 7(C_{20}-C_{11})^2 - 2(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{14(C_{20}- C_{11})^2+ 2(1-\rho^2)C_{20}C_{02}$$

$$+ 4C_{11}(C_{20}- C_{11}) - 2(2C_{30}- 3C_{21}+C_{12})\}$$

$$+ \frac{1}{(n')^2}\{- 7(C_{20}- C_{11})^2- 4C_{11}(C_{20}- C_{11})$$

$$+ 2(C_{30}- C_{21})\}] \quad . \qquad (2.93a)$$

Equation (2.93a) can be written as:

$$MSE(\bar{y}_{25}) - MSE(\bar{y}_{21}) = \bar{Y}^2(\frac{1}{n}-\frac{1}{n'})[2\{C_{30}(\frac{1}{n} - \frac{1}{n'}) + \frac{C_{12}}{n}$$

$$- (\frac{2}{n} - \frac{1}{n'})C_{21}\} - 7(C_{20}- C_{11})^2(\frac{1}{n} - \frac{1}{n'})$$

$$+ \frac{4C_{11}}{n'}(C_{20}- C_{11}) - \frac{2(1-\rho^2)}{n} C_{20}C_{02}] \qquad (2.93b)$$

It is easy to verify that if $C_{21} = C_{30} = C_{12}$ and $\rho < \frac{3(n-n')}{7n-3n'} \frac{C_x}{C_y}$ $(7n \neq 3n')$, then $\bar{y}_{25}$ will be better than $\bar{y}_{21}$. If $\rho > \frac{3(n-n')}{7n-3n'} \frac{C_x}{C_y}$, then $\bar{y}_{21}$ will be the more efficient estimator, provided, as before, $C_{21} = C_{30} = C_{12}$.

$$\begin{aligned} MSE(\bar{y}_{26}) - MSE(\bar{y}_{21}) = \bar{Y}^2 [& \frac{1}{n} \{2(C_{30} + C_{12} - 2C_{21}) \\ & - 5(C_{20} - C_{11})^2 - (1-\rho^2) C_{20} C_{02}\} \\ & + \frac{1}{nn'} \{2(3C_{21} - 2C_{30} - C_{12}) + 5(C_{20} - C_{11})^2 \\ & + (1-\rho^2) C_{20} C_{02} + 4C_{11}(C_{20} - C_{11})\} \\ & + \frac{1}{(n')^2} \{2(C_{30} - C_{21}) - 4C_{11}(C_{20} - C_{11})\}] \end{aligned} \qquad (2.94)$$

Again if $\rho < (1 - \frac{4n}{5n'}) \frac{C_x}{C_y}$ and $C_{21} = C_{30} = C_{12}$, then $\bar{y}_{26}$ is better than $\bar{y}_{21}$.

From (2.86) and (2.88),

$$MSE(\bar{y}_{26}) - MSE(\bar{y}_{25}) = \bar{Y}^2[\frac{1}{n^2}\{2(C_{20}-C_{11})^2+(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{-9(C_{20}-C_{11})^2-(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{(n')^2}\{7(C_{20}-C_{11})^2\}] \quad . \qquad (2.95)$$

If the preliminary sample is sufficiently large so that $n' > \frac{7}{2}n$, then irrespective of the value of ρ, $\bar{y}_{25}$ will be more efficient than $\bar{y}_{26}$.

From (2.84i) and (2.87),

$$MSE(\bar{y}_{23B}) - MSE(\bar{y}_{26}) = \bar{Y}^2[\frac{1}{n^2}\{2(C_{02}-2C_{11}+C_{20})$$

$$+ 2(2C_{21}-C_{12}-C_{30})+5(C_{20}-C_{11})^2$$

$$+ (1-\rho^2)C_{20}C_{02}+\frac{1}{\bar{Y}^2}(\sigma^2_{rX}-2\bar{X}\sigma_{rY})$$

$$+ \frac{2\sigma_{rX}}{\bar{Y}}(1+C_{20}-C_{11})\}$$

$$+ \frac{1}{nn'}\{2(2C_{30}-3C_{21}+C_{12}) - 5(C_{20}-C_{11})^2$$

$$- (1-\rho^2)C_{20}C_{02} - 4C_{11}(C_{20}-C_{11})$$

$$+ 2(4C_{11}-2C_{20}-C_{02}) - \frac{4\sigma_{rX}}{\bar{Y}}(1+C_{20}-C_{11})$$

$$- \frac{2}{\bar{Y}^2}(\sigma_{rX}^2+\bar{R}\sigma_{XY}-\bar{X}\sigma_{rY})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21}-C_{30}) + 4C_{11}(C_{20}-C_{11})$$

$$+ 2(C_{20}-2C_{11}) + \frac{2}{\bar{Y}}\sigma_{rX}(1+C_{20}-C_{11})$$

$$+ \frac{1}{\bar{Y}^2}(2\bar{R}\sigma_{XY}+\sigma_{rX}^2)\}] \quad . \qquad (2.96)$$

Using approximations (2.28) and (2.35e), we write (2.96) as:

$$MSE_a(\bar{y}_{23B})-MSE(\bar{y}_{26}) \doteq \bar{Y}^2\Big[\frac{1}{n^2}\{2(2C_{21}-C_{12}-C_{30})$$

$$+ 4(C_{20}-C_{11})^2+(1-\rho^2)C_{20}C_{02}\}$$

$$+ \frac{1}{nn'}\{2(2C_{30}-3C_{21}+C_{12})$$

$$- 3(C_{20}-C_{11})^2-(1-\rho^2)C_{20}C_{02}$$

$$- 6C_{11}(C_{20}-C_{11})\}$$

$$+ \frac{1}{(n')^2}\{2(C_{21}-C_{30})-(C_{20}-C_{11})^2$$

$$+ 6C_{11}(C_{20}-C_{11})\}\Big] \quad . \qquad (2.97)$$

It can easily be verified that if $C_{12} = C_{30} = C_{21}$ and $\rho \leq (\frac{4n'+n}{4n'+7n})\ \frac{C_x}{C_y}$, then $\bar{y}_{26}$ is superior to $\bar{y}_{23B}$ provided that approximations (2.28) and (2.35e) are valid.

Combining results (2.88a), (2.89b) and (2.90) - (2.97), we conclude that

1. If terms up to $O(\frac{1}{n'})$ only are considered then $\bar{y}_{2j}$, j = 1, 3B, 4, 5, 6 are equally efficient and each is more efficient than $\bar{y}_{22}$ if conditions (2.88c) are satisfied.

2. If terms up to $O(\frac{1}{(n')^2})$ are considered, it is not easy to compare $\bar{y}_{22}$ with the rest of the estimators. However, the five estimators can be ranked amongst themselves. If

(a) $$C_{21} = C_{30} = C_{12}$$

(b) $$n' > 3.5n$$

(c) $$\rho < \frac{5}{7} \frac{C_x}{C_y}$$

(d) approximations (2.28) and (2.35e) are valid, then

$$MSE(\bar{y}_{24}) = MSE(\bar{y}_{25}) \leq MSE(\bar{y}_{26}) < MSE_a(\bar{y}_{23B})$$

$$\leq MSE(\bar{y}_{21}) \quad . \tag{2.98}$$

We recall from (2.46) that this is almost identical with the set of inequalities we had for the corresponding estimators in Scheme I. We note that condition (d) above applies only to comparisons involving $MSE(\bar{y}_{23B})$.

We note further that for comparing the MSE's of $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$, we need only

(a) $$n' > 3.5n$$

and (b) $$\rho < \frac{C_x}{C_y}$$

for $$MSE(\bar{y}_{24}) = MSE(\bar{y}_{25}) < MSE(\bar{y}_{26}) .$$

We next compare the estimators under model (2.47). We recall that this implies a linear relationship between x_i and y_i. The relevant population parameters are given in (2.48).

After some straight-forward algebra, we obtain from (2.88a),

$$MSE(\bar{y}_{22}) - MSE(\bar{y}_{21}) = (\frac{1}{n} - \frac{1}{n'})[\alpha^2(\delta^2 - \frac{1}{\bar{X}^2})$$

$$+ \frac{1}{n}\{\alpha^2\gamma + a(\delta - \frac{1}{\bar{X}})\}\sigma_X^2$$

$$+ \alpha^2\{(\frac{1}{n} - \frac{1}{n'})(\delta\bar{X} - 1)^2\}$$

$$+ \alpha\{2C_{30}(\frac{\alpha}{n} - \frac{\alpha+\beta\bar{X}}{n'})$$

$$- C_{20}^2(\frac{9\alpha}{n} - \frac{6\alpha+4\beta\bar{X}}{n'})\}] . \qquad (2.99)$$

Under this model, if $\alpha > 0$, $\beta > 0$, $n' > (1+\frac{\beta\bar{X}}{\alpha})n$ and $\gamma_1 > \frac{9}{2}C_x$, then $\bar{y}_{21}$ is always superior to $\bar{y}_{22}$. We note further that under this model we can obtain specific conditions which ensure the superiority of $\bar{y}_{21}$ over $\bar{y}_{22}$ whereas tractable conditions of this type are not available in the general case. For n sufficiently large so that terms of $O(\frac{1}{n^2})$ can be neglected, $\bar{y}_{21}$ is always superior to $\bar{y}_{22}$. Again, if $\alpha = 0$ (i.e. the regression line passes through the origin), then $\bar{y}_{21}$ is always superior to $\bar{y}_{22}$.

Under model (2.47), we do not obtain similar tractable conditions from (2.83d). Even invoking approximations (2.28) and (2.35e) does not appear to lead to any major simplification. The same remarks apply to the comparison of $\bar{y}_{23B}$ with $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$ respectively.

Hence under model (2.51), we shall limit ourselves to comparing $\bar{y}_{21}$, $\bar{y}_{24}$, $\bar{y}_{25}$, and $\bar{y}_{26}$.

It is easy to verify from (2.86) and (2.92) that $\bar{y}_{24}$ and $\bar{y}_{25}$ are equally efficient under model (2.47). It is also easy to verify from (2.93) that $n' > \frac{7}{2}n$ is sufficient to ensure that $\bar{y}_{25}$ is better than $\bar{y}_{26}$. Finally,

$$MSE(\bar{y}_{26}) - MSE(\bar{y}_{21}) = (\frac{1}{n} - \frac{1}{n'})[2\alpha C_{30}(\frac{\alpha}{n} - \frac{\alpha+\beta\bar{X}}{n'})$$

$$+ C_{20}\{\frac{a\bar{X}}{n} - \alpha C_{20}(\frac{5\alpha}{n} - \frac{4\beta\bar{X}}{n'})\}] . \qquad (2.100)$$

Hence $\bar{y}_{26}$ is more efficient than $\bar{y}_{21}$ if

(a) $\alpha > 0 \ , \ \beta > 0$

(b) $C_{30} \leq 0$

(c) $n' > n(1 + \frac{\beta\bar{X}}{\alpha})$

and (d) $a\bar{X} < \alpha^2 C_{20}$

[The exact condition is

$$a\bar{X} < \alpha^2 C_{20}(1 + \frac{4\alpha}{\alpha+\beta\bar{X}})] \ .$$

Hence, if we consider the four estimators and if

(a) $\alpha > 0 \ , \ \beta > 0$

(b) $C_{30} \leq 0$

(c) $n' > n\max\{(1+\frac{\beta\bar{X}}{\alpha}) \ , \ \frac{7}{2}\}$

(d) $a\bar{X} < \alpha^2 C_{20}$

then

$$MSE(\bar{y}_{24}) = MSE(\bar{y}_{25}) \leq MSE(\bar{y}_{26}) \leq MSE(\bar{y}_{21}) .$$

We note that conditions (b) and (d) are not required when we want to rank the MSE's of $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$ only.

We now compare the MSE's of some of our estimators using the model given by 3(c) of Section C of this chapter. Under this model, (2.56), the regression of Y on X is assumed to be linear and X has a gamma distribution with parameter m. From (2.57) and (2.88a), we have:

$$MSE(\bar{y}_{22}) - MSE(\bar{y}_{21}) = (\frac{1}{n} - \frac{1}{n'})[\{\frac{(2m-1)\alpha^2}{m(m-1)^2}\}$$

$$+ \frac{\alpha^2}{m^2(m-1)}\{\frac{15m-3m^2-10}{n(m-2)} - \frac{4m-m^2-2}{n'(m-1)}$$

$$+ \frac{\delta(9m-2m^2-6)}{m(m-1)(m-2)}] . \qquad (2.101)$$

We note that for $2 < m < 3.69$, the expression on the r.h.s. of (2.101) is positive. Hence for m within this range, $\bar{y}_{21}$ is superior to $\bar{y}_{22}$. For $0 < m < 2$ $(m \neq 1, 2)$, or $m > 3.69$, the performance of $\bar{y}_{21}$ as compared to that of $\bar{y}_{22}$ depends on the values of n, n', α and δ.

For example, for $m = 4$, $\bar{y}_{21}$ will be superior to $\bar{y}_{22}$ if $\alpha^2\{28 + 3(\frac{1}{n} + \frac{1}{3n'})\} - 12\delta > 0$ where, as previously stated, $\delta = O(\frac{1}{n})$.

We note further that to $O(\frac{1}{n'})$, $\bar{y}_{22}$ is less efficient than $\bar{y}_{2j}$ $(j = 1, 3B, 4, 5, 6)$ if $m > \frac{1}{2}$.

From (2.57) and (2.92), we note, that irrespective of the model used, $\bar{y}_{24}$ and $\bar{y}_{25}$ are equally efficient. Again, from (2.61) and (2.95), $\bar{y}_{25}$ is superior to $\bar{y}_{26}$ if $n' > \frac{7}{2}n$.

From (2.57) and (2.96),

$$MSE(\bar{y}_{23B}) - MSE(\bar{y}_{26}) = -(\frac{1}{n} - \frac{1}{n'})[\frac{\alpha^2}{m^2(m-1)^2}\{\frac{2m^2 - 2m - 1}{n} - \frac{7m^2 - 12m + 4}{n'}\} + \frac{(m+1)\delta}{m(m-1)}] \qquad (2.102)$$

For $m > 1.5$ and $n' > \frac{7}{2}n$, the r.h.s. of (2.102) is negative and hence for these values $\bar{y}_{23B}$ is superior to $\bar{y}_{26}$. We note that approximations (2.28) and (2.35e) have not been invoked here, because for these approximations to be valid, we must assume

$$\frac{1}{m-1} \doteq \frac{1}{m} .$$

Such an assumption may not be reasonable for many populations encountered in practice.

From (2.57), (2.84i) and (2.85b),

$$MSE(\bar{y}_{23B}) - MSE(\bar{y}_{24}) = \frac{1}{n^2}\{\frac{-2\delta}{m(m-1)} + \frac{\alpha^2(6m^2-14m+9)}{m^2(m-1)^2} - \frac{6\alpha^2}{m^2}\}$$

$$+ \frac{1}{nn'}\{\frac{2\delta}{m(m-1)} + \frac{\alpha^2(-6m^2+16m-12)}{m^2(m-1)^2} + \frac{6\alpha^2}{m^2}\}$$

$$+ \frac{1}{(n')^2}\{\frac{\alpha^2(-2m+3)}{m^2(m-1)^2}\} \qquad (2.103)$$

It appears impossible to obtain simple conditions for the comparison of $\bar{y}_{23B}$ and $\bar{y}_{24}$. Hence in this section, we shall limit our ranking to $\bar{y}_{21}$, $\bar{y}_{22}$, $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$. To do this, we need, in addition to the above comparisons, the value of $MSE(\bar{y}_{26}) - MSE(\bar{y}_{21})$ under model (2.56). From (2.57) and (2.92),

$$MSE(\bar{y}_{26}) - MSE(\bar{y}_{21}) = -(\frac{1}{n} - \frac{1}{n'})[\frac{\alpha^2}{m^2}(\frac{1}{n} + \frac{4}{n'}) + \frac{8}{m}] \qquad (2.104)$$

Since $\delta > 0$, $\bar{y}_{26}$ is always preferable to $\bar{y}_{21}$ under this model.

To sum up, if

(a) $2 < m < 3.69$

(b) $n' > \frac{7}{2}n$

(c) $\alpha > 0 \ , \ \beta > 0$

then

$$MSE(\bar{y}_{24}) = MSE(\bar{y}_{25}) \leq MSE(\bar{y}_{26}) \leq MSE(\bar{y}_{21}) \leq MSE(\bar{y}_{22})$$

If we do not consider $\bar{y}_{22}$ in the ranking, then condition (a) above can be omitted and the relationships among the MSE's of $\bar{y}_{24}$, $\bar{y}_{25}$, $\bar{y}_{26}$ and $\bar{y}_{21}$ will be unchanged.

The final model we consider is (2.67). Under this model the regression of Y on X is quadratic and X is distributed as a normal variate with mean 1 and variance h. As in Section F, we restrict ourselves to a consideration of the relative magnitudes of the MSE's of $\bar{y}_{21}$, $\bar{y}_{24}$, $\bar{y}_{25}$ and $\bar{y}_{26}$.

We note again that, to $O(\frac{1}{(n')^2})$, $\bar{y}_{24}$ and $\bar{y}_{25}$ are equally efficient, regardless of the model used. Again from (2.90), $\bar{y}_{25}$ is more efficient than $\bar{y}_{26}$ if $n' > \frac{7}{2} n$. From (2.68) and (2.92),

$$MSE(\bar{y}_{26}) - MSE(\bar{y}_{21}) = \frac{1}{n^2}[-n\delta h - 5h^2\{\beta_0 + \beta_2(h-1)\}^2$$

$$- h^2\{8\beta_2(\beta_0 + \beta_2(h-1)) + 2\beta_2^2 h\}]$$

$$+ \frac{1}{nn'}[n\delta h + h^2\{5(\beta_0+\beta_2(h-1))^2$$

$$+ 4(5\beta_2+\beta_1)(\beta_0+\beta_2 h) - 12\beta_2^2 + 2\beta_2^2 h\}]$$

$$+ \frac{1}{(n')^2}[-4h^2\{\beta_0+\beta_2(h-1)\}\{\beta_1+3\beta_2\}$$

$$- 4\beta_2 h^2(\beta_1+2\beta_2)] \quad . \qquad (2.105a)$$

From (2.105a), it can be shown that if $n' > \frac{7}{2}n$ and $\beta_i > 0$ $(i = 0, 1, 2)$, $\bar{y}_{26}$ is superior to $\bar{y}_{21}$. It is difficult to derive other simple condition for this case.

From the preceding discussion, we conclude that for

(a) $$\beta_i > 0 \ , \quad i = 0, 2$$

(b) $$h > 1 - \frac{\beta_0}{\beta_2}$$

(c) $$n' > \frac{7}{2}n$$

$$MSE(\bar{y}_{24}) = MSE(\bar{y}_{25}) \leq MSE(\bar{y}_{26}) \leq MSE(\bar{y}_{21})$$

We note that if we put $\beta_2 = 0$, this is equivalent to our assuming that the regression of Y on X is linear. In this case, (2.105a) reduces to:

$$MSE(\bar{y}_{26}) - MSE(\bar{y}_{21}) = -(\frac{1}{n} - \frac{1}{n'})[\delta h + \frac{5\beta_0^2 h^2}{n} - \frac{4\beta_0\beta_1 h^2}{n'}] \qquad (2.105b)$$

Hence for $\beta_0 > 0$, and $4\beta_1 < 5\beta_0$, $\bar{y}_{26}$ is superior to $\bar{y}_{21}$. The original order of ranking of the four estimators is unchanged under the following amended conditions:

(a) $\beta_0 > 0$,

(b) $\beta_1 < 1.25\beta_0$,

(c) $n' > \frac{7}{2}n$.

H. Comparison between Independent (Scheme I) and Dependent (Scheme II) Estimators

We next consider the conditions under which Scheme I is more efficient than Scheme II. It may be recalled that under Scheme I the second phase sample is independent of the first phase sample while under Scheme II the second phase sample is a subsample.

As the computations involved in such comparisons are usually tedious and cumbersome and conclusions difficult to draw, we shall consider only

the classical ratio estimator. We shall also assume that the second phase sample size, n', is so large that it is sufficient to consider only terms of $O(\frac{1}{n'})$ in the MSE's. Since to $O(\frac{1}{n'})$, $MSE(\bar{y}_{j1}) = V(\bar{y}_{j1})$, $j = 1, 2$, it is sufficient to compare only the corresponding variances. To $O(\frac{1}{n'})$,

(1)
$$V(\bar{y}_{11}) = \frac{V_1}{n} + \frac{V_2}{n'} \tag{2.106}$$

where
$$V_1 = \sigma_Y^2 + R^2\sigma_X^2 - 2R\sigma_{XY}$$

and
$$V_2 = R^2\sigma_X^2 \quad .$$

(2)
$$V(\bar{y}_{21}) = \frac{V_1^*}{n} + \frac{V_2^*}{n'} \tag{2.107}$$

where
$$V_1^* = V_1$$

and
$$V_2^* = 2R\sigma_{XY} - R^2\sigma_X^2 \quad .$$

Hence, recalling that $R > 0$,

$$V(\bar{y}_{11}) < V(\bar{y}_{21})$$

if
$$\rho > \frac{C_x}{C_y} \quad . \tag{2.108}$$

As previously mentioned, Cochran [9] has stated that when X is the value of V at some previous time, the ratio $\frac{C_x}{C_y}$ may be approximately equal to 1. In this case, $V(\bar{y}_{11})$ cannot be less than $V(\bar{y}_{21})$, since $\rho \leq 1$.

However, it would not be very meaningful to compare the efficiencies of $\bar{y}_{11}$ and $\bar{y}_{21}$ without taking into account the costs of selection and measurement of the sampling units under the two schemes. For Scheme I, we consider the following simple cost function

$$C_0 = n(c_1 + c_2) + n'c_1 \tag{2.109}$$

where c_1 = cost per unit of measuring the X characteristic

c_2 = cost per unit of measuring the Y characteristic .

In this simple function, we have excluded the common fixed overhead costs. It is also implied in (2.109) that any unit in the first phase sample which appears again in the second phase sample has its X characteristic measured again. For the same total cost, C_0, the cost function under Scheme II is given by

$$C_0 = n^* c_2 + n^{*\prime} c_1 \tag{2.110}$$

where $n^{*\prime}$ = size of the preliminary sample

n^{*} = size of the second phase sample

and c_1, c_2 are as defined before.

In both expressions (2.109) and (2.110), we assume that

$$c_1 << c_2 ,$$

but c_1 is not negligible. Applying well-known results, we have:

$$V_{min}(\bar{y}_{11}) = \frac{[\sqrt{V_1}(c_1+c_2) + \sqrt{V_2}c_1]^2}{C_0} \qquad (2.111)$$

and

$$V_{min}(\bar{y}_{21}) = \frac{[\sqrt{V_1^*}c_2 + \sqrt{V_2^*}c_1]^2}{C_0} \quad . \qquad (2.112)$$

From (2.11) and (2.112)

$$V_{min}(\bar{y}_{11}) < V_{min}(\bar{y}_{21})$$

if

$$\sqrt{V_1^*}c_2 + \sqrt{V_2^*}c_1 > \sqrt{V_1}(c_1+c_2) + \sqrt{V_2}c_1 \quad . \qquad (2.113)$$

To illustrate clearly what the inequality (2.113) implies, consider the following artificial numerical example:

$$\bar{Y} = 6.9\ ,\quad \bar{X} = 5.4\ ,\quad \sigma_X^2 = 26.0\ ,\quad \sigma_Y^2 = 87.0 \quad \text{and} \quad \rho = 0.8$$

Thus

$$V_1^* = V_1 = 32.21$$

$$V_2 = 42.44$$

$$V_2^* = 54.79\ .$$

Let us also assume that $c_1 = \gamma c_2$, where by deductive reasoning,

$$0 < \gamma << 1\ .$$

It can easily be verified that the inequality (2.113) is satisfied if $\gamma < 0.102$.

We put $T = \dfrac{C_y}{C_x}$ and $\gamma = \dfrac{c_1}{c_2}$ (as before) and define

$$V_R = \frac{V_{min}(\bar{y}_{11})}{V_{min}(\bar{y}_{21})}$$

$$= \frac{\sqrt{(T^2 + 1 - 2\rho T)(1+\gamma)} + \sqrt{\gamma}}{\sqrt{T^2+1-2\rho T} + \sqrt{(2\rho T - 1)\gamma}}\ . \qquad (2.114)$$

The inequality (2.113) can now be written simply as

$$V_R < 1 \qquad (2.115)$$

We note that

(1) if $\rho \leq \frac{C_x}{C_y}$ (i.e. $\rho \leq T^{-1}$),

then $V_{min}(\bar{y}_{21})$ will be less than $V_{min}(\bar{y}_{11})$.

(2) If $\rho > \frac{C_x}{C_y}$,

then $V_R < 1$ if $\gamma < \frac{4\alpha^2}{(1-\alpha^2)^2}$, where

$$\alpha = \frac{\sqrt{2T\rho - 1} - 1}{\sqrt{T^2 + 1 - 2\rho T}} .$$

Table 2.1 gives the values of V_R for selected values of T, ρ and γ. It is known that for $T \leq 1$, $\bar{y}_{21}$ is always better than $\bar{y}_{11}$. However, we have included values of V_R for $T \leq 1$ to show how much worse $\bar{y}_{11}$ is. For $T = 1.5$, $\bar{y}_{21}$ continues to be superior to $\bar{y}_{11}$ except where ρ is high (i.e. 0.8 and above).

Table 2.1. Values of V_R for different values of T , ρ and γ

ρ	0.5		0.6		0.7		0.8		0.9	
T	γ=0.1	γ=0.01	γ=0.1	γ=0.01	γ=0.1	γ=0.01	γ=0.1	γ=0.01	γ=0.1	γ=0.01
0.8	-*	-*	-*	-*	1.03	1.09	1.23	1.08	1.20	1.07
1.0	1.04	1.00	1.21	1.06	1.16	1.05	1.12	1.04	1.08	1.02
1.5	1.10	1.03	1.06	1.02	1.03	1.00	0.99	0.99	0.95	0.97
2.0	1.04	1.01	1.03	1.00	0.98	0.98	0.95	0.97	0.92	0.96
2.5	1.01	1.00	0.99	0.99	0.96	0.98	0.94	0.97	0.91	0.96

*V_R undefined for these values.

Table 2.2. Values of γ_0 such that for $\gamma \leq \gamma_0$, $V_R < 1$

ρ					
T	0.5	0.6	0.7	0.8	0.9
1.5	-*	-*	0.01	0.17	0.97
1.75	-*	0.00	0.19	0.43	1.63
2.0	0.00	0.05	0.24	0.68	2.01
2.25	0.02	0.12	0.32	0.85	2.11
2.5	0.04	0.18	0.44	0.95	2.11

*V_R undefined for these values.

Table 2.2 shows the values of γ_0 such that for $\gamma \leq \gamma_0$, $V_R < 1$. It shows that if ρ is only slightly greater than $1/T$, then γ must be relatively small. However, when ρ is very much greater than $1/T$, γ could be greater than 1. If $\gamma_0 > 1$, this implies that any small γ will result in $V_R < 1$.

As a rule of thumb, it would not be inappropriate to suggest that the independent second phase sample estimator will generally be superior to $\bar{y}_{21}$ if

(1) $$\rho >> \frac{C_x}{C_y} ,$$

and

(2) the cost per unit of measuring the concomitant variable is very small compared with the cost per unit of measuring the variable of interest. It is not unrealistic to have $\gamma = 0.1$ (i.e. $c_1 = \frac{1}{10} c_2$), say, especially in cases where the values of the X variable can be obtained from files.

To the same order of approximation, we get identical results for all other pairs of estimators except $\bar{y}_{12}$ and $\bar{y}_{22}$. For this pair a similar type of comparison can be made by merely replacing T with $\bar{T}$ in the preceding arguments. $\bar{T}$ is given by

$$\bar{T} = RC_y/\bar{R}C_x .$$

III. THE REGRESSION ESTIMATOR

In this chapter, we shall obtain a more exact expression for the bias, variance and MSE of the estimator $\bar{y}_{Rd}$. As in Chapter II, all expressions will be given to $O(\frac{1}{(n')^2})$.

We recall from (1.22) that

$$E(\bar{y}_{Rd}) = \bar{Y} + \mathrm{Cov}(b, \bar{x}_{n'}) - \mathrm{Cov}(b, \bar{x}_n) \tag{3.1}$$

To simplify (3.1), we first evaluate $\mathrm{Cov}(b, \bar{x}_n)$, where $b = \frac{s_{xy}}{s_x^2}$. With the usual notation, we can write

$$s_{xy} = S_{XY}(1 + \delta s_{xy})$$

$$s_x^2 = S_X^2(1 + \delta s_x^2)$$

$$\bar{x}_n = \bar{X}(1 + \delta \bar{x}_n) \quad .$$

Using this approach, we can easily show that

$$\mathrm{Cov}(b, \bar{x}_n) = \beta \bar{X}[\frac{1}{n}(\frac{C_{21}}{C_{11}} - \frac{C_{30}}{C_{20}})$$

$$+\frac{1}{n^2}\left(\frac{C_{50}}{C_{20}^2}+\frac{C_{40}C_{21}}{C_{11}C_{20}^2}+\frac{2C_{31}C_{30}}{C_{11}C_{20}^2}+\frac{3C_{21}}{C_{11}}-\frac{3C_{40}C_{30}}{C_{20}^3}\right.$$

$$\left.-\frac{C_{41}}{C_{11}C_{20}}-\frac{3C_{30}}{C_{20}}\right)] \qquad (3.2)$$

This agrees with the result obtained by Fuller and Johnson [17]. We now put

$\bar{y}_{1R} = \bar{y}_{Rd}$ in Scheme I as defined in Chapter II

$\bar{y}_{2R} = \bar{y}_{Rd}$ in Scheme II

Under Scheme I where the second phase sample is independent of the first phase sample, b and $\bar{x}_n$, are independent and hence

$$E(\bar{y}_{1R}) = \bar{Y}[1+\frac{\beta}{R}\{\frac{1}{n}(\frac{C_{30}}{C_{20}}-\frac{C_{21}}{C_{11}})$$

$$+\frac{1}{n^2}[3(\frac{C_{30}}{C_{20}}-\frac{C_{21}}{C_{11}})+\frac{1}{C_{20}}(\frac{C_{41}}{C_{11}}-\frac{C_{50}}{C_{20}})$$

$$+\frac{C_{40}}{C_{20}^2}(\frac{C_{30}}{C_{20}}-\frac{C_{21}}{C_{11}})+2\frac{C_{30}}{C_{20}^2}(\frac{C_{40}}{C_{20}}-\frac{C_{31}}{C_{11}})]\}] \qquad (3.3)$$

We recall from (2.16) that in the case of Scheme I $E(\delta\bar{x}_{n'}|x_1, x_2, \ldots x_n, y_1, y_2, \ldots y_n) \doteq \frac{n}{n'}\delta\bar{x}_n$ for N large. Hence the second term of expression (3.1) reduces to

$$\mathrm{Cov}(b, \bar{x}_{n'}) = \frac{n}{n'}\mathrm{Cov}(b, \bar{x}_n) \tag{3.4}$$

Combining equations (3.1), (3.2) and (3.4), we have:

$$E(\bar{y}_{2R}) = \bar{Y}[1 + \frac{\beta}{R}\{(\frac{1}{n} - \frac{1}{n'})(\frac{C_{30}}{C_{20}} - \frac{C_{21}}{C_{11}})$$

$$+ \frac{1}{n}(\frac{1}{n} - \frac{1}{n'})[3(\frac{C_{30}}{C_{20}} - \frac{C_{21}}{C_{11}}) + \frac{1}{C_{20}}(\frac{C_{41}}{C_{11}} - \frac{C_{50}}{C_{20}})$$

$$+ \frac{C_{40}}{C_{20}^2}(\frac{C_{30}}{C_{20}} - \frac{C_{21}}{C_{11}}) + 2\frac{C_{30}}{C_{20}^2}(\frac{C_{40}}{C_{20}} - \frac{C_{31}}{C_{11}})]\}] \tag{3.5}$$

It is easy to verify that where (X, Y) follow the bivariate normal distribution $E(\bar{y}_{1R}) = E(\bar{y}_{2R}) = \bar{Y}$, since all odd order product moment coefficients (C_{ij}'s) are zero.

We follow the same procedure to derive expressions for $V(\bar{y}_{1R})$ and $V(\bar{y}_{2R})$ to $O(\frac{1}{(n')^2})$. We recall again from (2.17) that under Scheme I,

$$E[(\delta\bar{x}_{n'} - \delta\bar{x}_n)^2 | x_1, x_2, \ldots, x_n, y_1, y_2, \ldots y_n]$$

$$= \delta^2\bar{x}_{n'} + \delta^2\bar{x}_n \quad . \tag{3.6}$$

Using (3.6), we obtain, after some simplification

$$Eb^2(\bar{x}_{n'} - \bar{x}_n)^2 = \beta^2\bar{X}^2[(\frac{1}{n} + \frac{1}{n'})C_{20}$$

$$+ \frac{1}{n^2}(\frac{C_{40}}{C_{20}} + \frac{C_{20}C_{22}}{C_{11}^2} + 6\frac{C_{30}^2}{C_{20}^2} + 2\frac{C_{21}^2}{C_{11}^2}$$

$$- \frac{8C_{30}C_{21}}{C_{11}C_{20}} - 2\frac{C_{31}}{C_{11}})$$

$$+ \frac{1}{nn'}(\frac{C_{22}C_{20}}{C_{11}^2} - 4\frac{C_{31}}{C_{11}} + 3\frac{C_{40}}{C_{20}})] \, . \tag{3.7}$$

and

$$E\bar{y}_n b(\bar{x}_{n'} - \bar{x}_n) = -\beta\bar{X}\bar{Y}[\frac{1}{n}(\frac{C_{21}}{C_{11}} + C_{11} - \frac{C_{30}}{C_{20}})$$

$$+ \frac{1}{n^2}\{C_{11} + \frac{1}{C_{11}}(C_{22} - C_{20}C_{02} + 3C_{21})$$

$$- \frac{1}{C_{20}}(3C_{30}+2C_{31})$$

$$- \frac{1}{C_{20}C_{11}}(C_{41}+C_{21}^2+C_{30}C_{12})$$

$$+ \frac{1}{C_{20}^2}(C_{50}+C_{40}C_{11}+2C_{30}C_{21}) + \frac{3C_{40}C_{30}}{C_{20}^3}$$

$$+ \frac{1}{C_{20}^2 C_{11}}(C_{40}C_{21}+2C_{31}C_{30})\}] \qquad (3.8)$$

Combining equations (3.5), ..., (3.8), we obtain:

$$V(\bar{y}_{1R}) = \bar{Y}^2[\frac{1}{n}(C_{02}-\frac{C_{11}^2}{C_{20}}) + \frac{1}{n'}\frac{C_{11}^2}{C_{20}}$$

$$+ \frac{1}{n^2}\{2(C_{02}-\frac{C_{11}^2}{C_{20}}) - \frac{C_{22}}{C_{20}}$$

$$+ \frac{1}{C_{20}^2}(2C_{31}C_{11}+3C_{21}^2+2C_{30}C_{12})$$

$$- \frac{C_{11}}{C_{20}^3}(10C_{30}C_{21} + C_{40}C_{11}) + \frac{5C_{30}^2C_{11}^2}{C_{20}^4}$$

$$+ \frac{1}{nn'}\{\frac{1}{C_{20}}(C_{22} - \frac{4C_{31}C_{11}}{C_{20}} + \frac{3C_{40}C_{11}^2}{C_{20}^2}\}] \qquad (3.9)$$

We obtain $V(\bar{y}_{2R})$ in a similar way by noting that for dependent second phase sampling (Scheme II), with large N,

(1) $$E[(\delta\bar{x}_{n'} - \delta\bar{x}_n)|x_1, x_2, \ldots, x_n, y_1, y_2, \ldots y_n]$$

$$\doteq -(1 - \frac{n}{n'})\delta\bar{x}_n \qquad (3.10a)$$

and

(2) $$E[(\delta\bar{x}_{n'} - \delta\bar{x}_n)^2|x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n]$$

$$\doteq (1 - \frac{n}{n'})^2\delta^2\bar{x}_n + (\frac{1}{n'} - \frac{n}{(n')^2})C_{20} \qquad (3.10b)$$

Hence, under Scheme II,

$$V(\bar{y}_{2R}) = \bar{Y}^2[\frac{1}{n}(C_{02} - \frac{C_{11}^2}{C_{20}}) + \frac{1}{n'}\frac{C_{11}^2}{C_{20}}$$

$$+\frac{1}{n^2}\{2(C_{02}-\frac{C_{11}^2}{C_{20}})-\frac{C_{22}}{C_{20}}$$

$$+\frac{1}{C_{20}^2}(2C_{31}C_{11}+3C_{21}^2+2C_{30}C_{12})$$

$$-\frac{C_{11}}{C_{20}^3}(10C_{30}C_{21}+C_{40}C_{11})+\frac{5C_{30}^2C_{11}^2}{C_{20}^4}\}$$

$$+\frac{1}{nn'}\{-2(C_{02}-\frac{C_{11}^2}{C_{20}})+\frac{C_{22}}{C_{20}}$$

$$-\frac{2}{C_{20}^2}(2C_{31}C_{11}+C_{30}C_{12}+2C_{21}^2)$$

$$+\frac{C_{11}}{C_{20}^3}(3C_{40}C_{11}+16C_{30}C_{21})-\frac{10C_{30}^2C_{11}^2}{C_{20}^4}\}$$

$$+\frac{1}{(n')^2}\{\frac{1}{C_{20}^2}(C_{21}^2+2C_{31}C_{11})$$

$$-\frac{2C_{11}}{C_{20}^3}(3C_{30}C_{21}+C_{40}C_{11})$$

$$+\frac{5C_{30}^2C_{11}^2}{C_{20}^4}\}] \qquad (3.11)$$

From (3.3) and (3.9),

$$MSE(\bar{y}_{1R}) = \bar{Y}^2[\frac{1}{n}(C_{02} - \frac{C_{11}^2}{C_{20}}) + \frac{1}{n'}\frac{C_{11}^2}{C_{20}}$$

$$+ \frac{1}{n^2}\{2(C_{02} - \frac{C_{11}^2}{C_{20}}) - \frac{C_{22}}{C_{20}}$$

$$+ \frac{2}{C_{20}^2}(C_{31}C_{11} + 2C_{21}^2 + C_{30}C_{12})$$

$$- \frac{C_{11}}{C_{20}^3}(12C_{30}C_{21} + C_{40}C_{11}) + \frac{6C_{30}^2C_{11}^3}{C_{20}^4}\}$$

$$+ \frac{1}{nn'}\{\frac{1}{C_{20}}(C_{22} - \frac{4C_{31}C_{11}}{C_{20}} + \frac{3C_{40}C_{11}^2}{C_{20}^2})\}] \quad (3.12)$$

From (3.5) and (3.11),

$$MSE(\bar{y}_{2R}) = \bar{Y}^2[\frac{1}{n}(C_{02} - \frac{C_{11}^2}{C_{20}}) + \frac{1}{n'}\frac{C_{11}^2}{C_{20}}$$

$$+ \frac{1}{n^2}\{2(C_{02} - \frac{C_{11}^2}{C_{20}}) - \frac{C_{22}}{C_{20}}$$

$$+ \frac{2}{C_{20}^2}(C_{31}C_{11} + 2C_{21}^2 + C_{30}C_{12})$$

$$- \frac{C_{11}}{C_{20}^3}(12C_{30}C_{21}+C_{40}C_{11}) + \frac{6C_{30}^2C_{11}^2}{C_{20}^4}$$

$$+ \frac{1}{nn'}\{-2(C_{02}- \frac{C_{11}^2}{C_{20}}) + \frac{C_{22}}{C_{20}}$$

$$- \frac{2}{C_{20}^2}(2C_{31}C_{11}+C_{30}C_{12}+3C_{21}^2)$$

$$+ \frac{C_{11}}{C_{20}^3}(3C_{40}C_{11}+20C_{30}C_{21}) - \frac{12C_{30}^2C_{11}^2}{C_{20}^4}\}$$

$$+ \frac{1}{(n')^2}\{\frac{2}{C_{20}^2}(C_{21}^2+C_{31}C_{11})$$

$$- \frac{2C_{11}}{C_{20}^3}(4C_{30}C_{21}+C_{40}C_{11})$$

$$+ \frac{6C_{30}^2C_{11}^2}{C_{20}^4}\}] \qquad (3.13)$$

We note that unlike the ratio-type estimators considered in Chapter II, the two-phase regression estimators for both Schemes I and II have identical MSE's if terms up to $O(\frac{1}{n^2})$ only are considered.

IV. SRIVASTAVA'S ESTIMATOR

In Chapter I, mention was made of Srivastava's estimator

$$\bar{y}_{\alpha} = \bar{y}_n \left(\frac{\bar{x}_n}{\bar{X}}\right)^{\alpha} \tag{4.1}$$

As indicated earlier, if we put $\alpha = 1$, we obtain the usual product estimator

$$\bar{y}_p^{*} = \bar{y}_n \left(\frac{\bar{x}_n}{\bar{X}}\right) \tag{4.2}$$

The two-phase sample equivalent of the product estimator is

$$\bar{y}_p = \bar{y}_n \left(\frac{\bar{x}_n}{\bar{x}_{n'}}\right) \tag{4.3}$$

Under Scheme I of Chapter II, where the preliminary sample of size n' is independent of the second phase sample of size n,

$$\bar{y}_{1p} = \bar{y}_n \left(\frac{\bar{x}_n}{\bar{x}_{n'}}\right) \tag{4.4a}$$

$$E(\bar{y}_{1p}) = \bar{Y}[1 + \frac{1}{n} C_{11} + \frac{1}{n'} C_{20}$$

$$+ \frac{1}{n}(\frac{C_{20}C_{11}}{n} + \frac{3C_{20}^2}{n'} - \frac{C_{30}}{n'})] . \qquad (4.4b)$$

$$V(\bar{y}_{1p}) = \bar{Y}^2[\frac{1}{n}(C_{20}+2C_{11}+C_{02}) + \frac{1}{n'} C_{20}$$

$$+ \frac{1}{n^2}(2C_{12}+2C_{21}+C_{11}^2+C_{20}C_{02})$$

$$+ \frac{1}{nn'}(8C_{20}C_{11}+3C_{20}^2+3C_{20}C_{02})$$

$$+ \frac{1}{(n')^2}(8C_{20}^2-2C_{30})] . \qquad (4.4c)$$

From equations (4.4b) and (4.4c),

$$MSE(\bar{y}_{1p}) = \bar{Y}^2[\frac{1}{n}(C_{02}+2C_{11}+C_{20}) + \frac{1}{n'}C_{20}$$

$$+ \frac{1}{n^2}(2C_{12}+2C_{21}+2C_{11}^2+C_{20}C_{02})$$

$$+ \frac{C_{20}}{nn'}(10C_{11}+3C_{20}+3C_{02})$$

$$+ \frac{1}{(n')^2}(9C_{20}^2-2C_{30})] \quad . \tag{4.4d}$$

Under Scheme II of Chapter II where the second phase sample is a subsample of the first phase sample,

$$\bar{y}_{2p} = \bar{y}_n(\frac{\bar{x}_n}{\bar{x}_{n'}}) \tag{4.5a}$$

$$E(\bar{y}_{2p}) = \bar{Y}[1 + (\frac{1}{n}-\frac{1}{n'})(C_{11}-\frac{C_{21}}{n'}+\frac{C_{20}C_{11}}{n'})] \quad . \tag{4.5b}$$

$$V(\bar{y}_{2p}) = \bar{Y}^2[\frac{1}{n}(C_{02}+2C_{11}+C_{20}) - \frac{1}{n'}(C_{20}+2C_{11})$$

$$+ \frac{1}{n^2}(2C_{12}+2C_{21}+C_{11}^2+C_{20}C_{02})$$

$$+ \frac{1}{nn'}(3C_{20}^2-C_{20}C_{02}-6C_{11}^2-2C_{12}$$

$$- 6C_{21}-2C_{30}-2C_{20}C_{11})$$

$$+ \frac{1}{(n')^2}(4C_{21} + 5C_{11}^2 + 2C_{30} - 3C_{20}^2$$

$$+ 2C_{20}C_{11})] \ . \qquad (4.5c)$$

From equations (4.5b) and (4.5c),

$$MSE(\bar{y}_{2p}) = \bar{Y}^2[\frac{1}{n}(C_{02} + 2C_{11} + C_{20}) - \frac{1}{n'}(C_{20} + 2C_{11})$$

$$+ \frac{1}{n^2}(2C_{12} + 2C_{21} + 2C_{11}^2 + C_{20}C_{02})$$

$$+ \frac{1}{nn'}\{C_{20}(3C_{20} - 2C_{11} - C_{02})$$

$$- 2(C_{12} + 3C_{21} + C_{30}) - 8C_{11}^2\}$$

$$+ \frac{1}{(n')^2}\{C_{20}(2C_{11} - 3C_{20})$$

$$+ 2(2C_{21} + C_{30}) + 6C_{11}^2\}] \ . \qquad (4.5d)$$

More generally, Srivastava [53] has shown that for $\bar{y}_\alpha$ to have minimum variance in single phase sampling, optimum $\alpha = -\frac{C_{11}}{C_{20}}$, where only terms to $O(\frac{1}{n})$ are considered. Hereafter, we shall refer to optimum α simply as $\alpha(= -C_{11}/C_{20})$. Srivastava has suggested that α could be estimated from a pilot sample survey or past data or experience. We are concerned here with only two methods of estimating α:

(1) We could obtain an estimate $\hat{\alpha}$ from a preliminary simple random sample of size m, where for convenience we assume $m < n < m^2$, n being the size of the second phase (or actual) survey. An obvious choice of $\hat{\alpha}$ is

$$\hat{\alpha}_m = -\frac{\bar{x}_m s_{xy}(m)}{\bar{y}_m s_x^2(m)} \tag{4.6a}$$

where $\bar{x}_m$, $\bar{y}_m$, $s_x^2(m)$, $s_{xy}(m)$ are the usual sample values computed from the preliminary sample of size m.

(2) Secondly, we could estimate α from the single phase sample of size n in which case

$$\hat{\alpha}_n = -\frac{\bar{x}_n s_{xy}}{\bar{y}_n s_x^2} \tag{4.6b}$$

where $\bar{x}_n$, $\bar{y}_n$, s_{xy}, s_x^2 are calculated from the sample of size n.

We note that we can write the estimator (4.1) as

$$\bar{y}_{\hat{\alpha}} = \bar{Y}(1 + \delta\bar{y}_n)(1 + \delta\bar{x}_n)^{\hat{\alpha}} \qquad (4.7a)$$

For given $\hat{\alpha}$, expression (4.7a) can be expanded by the application of the binomial theorem for any index. More simply, we can apply Taylor's expansion and write

$$\bar{y}_{\hat{\alpha}} = \bar{Y}(1 + \delta\bar{y}_n) f(\delta\bar{x}_n, \hat{\alpha}) , \qquad (4.7b)$$

where $$f(\delta\bar{x}_n, \hat{\alpha}) = (1 + \delta\bar{x}_n)^{\hat{\alpha}} .$$

Expanding $f(\delta\bar{x}_n, \hat{\alpha})$ as a Taylor series in the neighborhood of $(0, \alpha)$, we have:

$$f(\delta\bar{x}_n, \hat{\alpha}) \doteq 1 + \hat{\alpha}\,\delta\bar{x}_n + \frac{\hat{\alpha}(\hat{\alpha}-1)}{2}\,\delta^2\bar{x}_n + \left\{\frac{\alpha^2(3-2\alpha) + \hat{\alpha}(3\alpha^2 - 6\alpha + 2)}{6}\right\}\delta^3\bar{x}_n + \frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)}{24}\,\delta^4\bar{x}_n . \qquad (4.7c)$$

Hence, to $O(\frac{1}{n^2})$, we can write (4.7b) as

$$\bar{y}_\alpha \doteq \bar{Y}[1+\hat{\alpha}\,\delta\bar{x}_n+\delta\bar{y}_n+\frac{\hat{\alpha}(\hat{\alpha}-1)}{2}\delta^2\bar{x}_n+\hat{\alpha}\,\delta\bar{x}_n\delta\bar{y}_n$$

$$+\{\frac{\alpha^2(3-2\alpha)+\hat{\alpha}(3\alpha^2-6\alpha+2)}{6}\}\delta^3\bar{x}_n+\frac{\hat{\alpha}(\hat{\alpha}-1)}{2}\delta^2\bar{x}_n\delta\bar{y}_n$$

$$+\frac{\alpha(\alpha-1)(\alpha-2)(\alpha-3)}{24}\,\delta^4\bar{x}_n$$

$$+\{\frac{\alpha^2(3-2\alpha)+\hat{\alpha}(3\alpha^2-6\alpha+2)}{6}\}\,\delta^3\bar{x}_n\delta\bar{y}_n] \,. \qquad (4.8)$$

We recall that $\alpha = -\,C_{11}/C_{20}$ and hence

(1) $$E(\hat{\alpha}_m) = \alpha[1+\frac{1}{m}\{(C_{02}+\alpha C_{20})+\frac{1}{C_{20}}[(\alpha+1)C_{12}-\alpha C_{30}-C_{21}]$$

$$+\frac{1}{\alpha C_{20}^2}(\alpha C_{40}+C_{31})\}] \qquad (4.9a)$$

(2) $$E(\hat{\alpha}_m^2) = \alpha^2[1+\frac{1}{m}\{(3C_{02}+4\alpha C_{20}+C_{20})$$

$$+ \frac{4}{C_{20}}\left[\frac{1}{\alpha}(C_{12} - C_{21}) + C_{21} - C_{30}\right]$$

$$+ \frac{1}{C_{20}^2}\left(\frac{4C_{31}}{\alpha} + 3C_{40} + \frac{C_{22}}{\alpha^2}\right)\}\Big] \; . \qquad (4.9b)$$

From estimators (4.1) and (4.6a), we define the estimator in the case of the independent preliminary sample as

$$\bar{y}_{1\alpha} = \bar{y}_n\left(\frac{\bar{x}_n}{\bar{X}}\right)^{\hat{\alpha}_m} \qquad (4.10)$$

Hence, from equations (4.8) - (4.10),

$$E(\bar{y}_{1\alpha}) = \bar{Y}\left[1 - \frac{\alpha(\alpha+1)}{2n}C_{20}\right.$$

$$+ \frac{1}{2mn}\{\alpha C_{20}[(\alpha-1)C_{02} + 2\alpha^2 C_{20}]$$

$$+ \alpha(\alpha+1)(C_{21} + C_{30} - 2C_{12}) + C_{21} - C_{12}$$

$$\left. + \frac{1}{C_{20}}[(2\alpha-1)C_{31} + \alpha(\alpha-1)C_{40} + C_{22}]\}\right]$$

$$+ \frac{\alpha(\alpha-1)}{24n^2}\{4[(\alpha-2)C_{30}+3C_{21}]$$

$$- 9(\alpha+1)(\alpha-2)C_{20}^2\}] \quad . \qquad (4.11)$$

We note that

$$MSE(\bar{y}_1) = E(\bar{y}_{1\alpha}-\bar{Y})^2 \quad .$$

Hence from equations (4.8), (4.9a) and (4.9b), we obtain after some straightforward algebra,

$$MSE(\bar{y}_{1\alpha}) = \bar{Y}^2[\frac{1}{n}(C_{02}-\alpha^2 C_{20})+\frac{1}{nm}\{\alpha^2 C_{20}(C_{02}+C_{20}+2\ C_{20})$$

$$+ 2\alpha[(3\alpha-2)C_{21}+\alpha(\alpha-2)C_{30}-(\alpha+2)(\alpha-1)C_{12}]$$

$$+ \frac{1}{C_{20}}(2\alpha C_{31}+\alpha^2 C_{40}+C_{22})\}$$

$$+\frac{1}{n^2}\{\alpha(2\alpha-1)C_{20}C_{02}+\alpha[\alpha(\alpha-1)C_{30}+(3\alpha-1)C_{21}$$

$$+2C_{12}]+\frac{\alpha^2}{4}(3+10\alpha-5\alpha^2)C_{20}^2\}] \quad . \tag{4.12}$$

We note that if n and m $(m < n)$ are sufficiently large so that terms of $O(\frac{1}{nm})$ are negligible, then (4.12) reduces to

$$MSE(\bar{y}_{1\alpha}) = \frac{\bar{Y}^2}{n}(1-\rho^2)\,C_{02} \tag{4.13}$$

which is the MSE of the usual regression estimator. The similarity between Srivastava's estimator and the regression estimator will be examined in greater detail later.

We next consider the scheme in which the estimate of α is obtained from the actual survey itself. From equations (4.1) and (4.6b), we define

$$\bar{y}_{2\alpha} = \bar{y}_n\left(\frac{\bar{x}_n}{\bar{X}}\right)^{\hat{\alpha}_n} \tag{4.14}$$

It can be shown that under this scheme,

(1) $$E(\hat{\alpha}_n - \alpha)\bar{\delta x}_n = \frac{1}{n}[\alpha(1+\alpha)C_{20} - \frac{1}{C_{20}}(C_{21} + \alpha C_{30})]$$

$$+ \frac{1}{n^2}[\alpha C_{20}\{\alpha(1 + 2\alpha C_{20}) + (1 + 3\alpha)C_{02}\}$$

$$+ 2\alpha(C_{12} - C_{21}) - 2\alpha^2(C_{30} - C_{21}) - C_{02}$$

$$+ \frac{1}{C_{20}}\{C_{22} + \alpha^2 C_{40} + 2\alpha C_{31}$$

$$- (3 + C_{02})(C_{21} + \alpha C_{30})\}$$

$$+ \frac{1}{C_{20}^2}\{\alpha C_{50} + C_{41}$$

$$+ (2\alpha C_{30} + C_{21})(C_{30} - C_{21}) + C_{30}(C_{21} - C_{12})\}$$

$$- \frac{1}{C_{20}^3}\{C_{40}(C_{21} + \alpha C_{30})$$

$$+ 2C_{30}(C_{31} + \alpha C_{40})\}] \qquad (4.15a)$$

(2) $$E(\hat{\alpha}_n - \alpha)\delta^2\bar{x}_n = \frac{1}{n^2}[\alpha C_{20}\{\alpha(2\alpha+3)C_{20} + C_{02}\}$$

$$+ \{C_{12} + \alpha C_{21} - 2\alpha(1+\alpha)C_{30}\}$$

$$+ \frac{2C_{30}}{C_{20}^2}(\alpha C_{30} + C_{21})] \qquad (4.15b)$$

(3) $$E(\hat{\alpha}_n - \alpha)\delta\bar{x}_n\delta\bar{y}_n = \frac{1}{n^2}[C_{02} - \alpha C_{20}\{\alpha C_{20}(3+2\alpha) + \alpha + C_{02}\}$$

$$+ 2\alpha C_{21}(1-\alpha) - (3\alpha+1)C_{12} + 2\alpha^2 C_{30}$$

$$+ \frac{1}{C_{20}}\{C_{02}(\alpha C_{30} + C_{21}) - (\alpha^2 C_{40} + 2\alpha C_{31} + C_{22})\}$$

$$+ \frac{1}{C_{20}^2}\{C_{21}(\alpha C_{30} + C_{21})$$

$$+ C_{30}(\alpha C_{30} + C_{12})\}] \qquad (4.15c)$$

(4) $$E(\hat{\alpha}_n - \alpha)\delta^3\bar{x}_n = \frac{3\alpha}{n^2}[C_{20}^2(1+\alpha) + (C_{21} - C_{30})] \qquad (4.15d)$$

$$(5)\qquad E(\hat{\alpha}_n - \alpha)\delta^2\bar{x}_n\delta\bar{y}_n = \frac{1}{n^2}[\alpha(2\alpha C_{30}+C_{21}) - C_{12}$$

$$- \alpha C_{20}\{\alpha C_{20}(3+2\alpha) + C_{02}\}] \qquad (4.15e)$$

$$(6)\qquad E(\hat{\alpha}_n - \alpha)^2\delta^2\bar{x}_n = \frac{1}{n^2}[\alpha^2 C_{20}\{C_{20}(3+6\alpha+2\alpha^2) + C_{02}\}$$

$$+ \alpha\{2(2\alpha C_{21}+C_{12}) - 2\alpha(3+2\alpha)C_{30}\}$$

$$+ \frac{1}{C_{20}}\{\alpha^2 C_{40}+2\alpha C_{31}+C_{22}\}$$

$$+ \frac{1}{C_{20}^2}\{2(C_{21}+\alpha C_{30})^2\}] \qquad (4.15f)$$

$$(7)\qquad E(\hat{\alpha}_n - \alpha)\delta\bar{x}_n\delta^2\bar{y}_n = \frac{1}{n^2}[\alpha C_{20}\{2\alpha^2 C_{20}+(1+3\alpha)C_{02}\}$$

$$+ 2\alpha(C_{12}+\alpha C_{21}) - \frac{C_{02}}{C_{20}}(C_{21}+\alpha C_{30})] \qquad (4.15g)$$

Hence from (4.8) and (4.15a) - (4.15g), we obtain after some simplification:

$$
\begin{aligned}
E(\bar{y}_{2\alpha}) \doteq \bar{Y}\Big[1 + \frac{1}{n}\Big\{\frac{\alpha(1+\alpha)}{2} C_{20} - \frac{1}{C_{20}}(C_{21} + \alpha C_{30})\Big\}\Big.
\\
+ \frac{1}{n^2}\Big\{\frac{\alpha}{8} C_{20}\big[C_{20}(2 - 25\alpha + 18\alpha^2 + 17\alpha^3) + 28\alpha C_{02}\big]
\\
+ \frac{1}{2}\big[2(\alpha-1)C_{12} + (12\alpha^2 - 9\alpha + 2)C_{21}
\\
- (4\alpha^3 + 13\alpha^2 - 8\alpha + 2)C_{30}\big]
\\
+ \frac{1}{2C_{20}}\big[(2\alpha C_{31} + \alpha^2 C_{40} + C_{22}) - 6(C_{21} + \alpha C_{30})\big]
\\
+ \frac{1}{C_{20}^2}\big[(\alpha C_{50} + C_{41}) + (C_{21} + \alpha C_{30})^2
\\
+ (1+\alpha)(2\alpha C_{30} + C_{21})C_{30}\big]
\\
\Big.\Big. - \frac{1}{C_{20}^3}\big[2C_{30}(C_{31} + \alpha C_{40}) + C_{40}(C_{21} + \alpha C_{30})\big]\Big\}\Big] \; . \qquad (4.16)
\end{aligned}
$$

We note that in terms of the regression coefficient β mentioned in Chapter III,

$$\alpha\bar{Y} = -\beta\bar{X} \quad . \tag{4.17a}$$

Hence to $O(\frac{1}{n})$, the bias of $\bar{y}_{2\alpha}$ can be written as:

$$\beta(\bar{y}_{2\alpha}) = -\frac{\beta\bar{X}}{n}\left[\frac{(C_{20}-C_{11})}{2} + \left(\frac{C_{21}}{C_{11}} - \frac{C_{30}}{C_{20}}\right)\right] \tag{4.17b}$$

When the first term on the r.h.s. of (4.17b) is omitted, the remaining expression is exactly the bias, to $O(\frac{1}{n})$, of the regression estimator in single phase sampling. Some explanation for the presence of the first term in (4.17b) in relation to the bias of the regression estimator will be given later in this chapter.

Again using equations (4.8) and (4.15a) - (4.15g), we obtain after some rather tedious algebra:

$$MSE(\bar{y}_{2\alpha}) = \bar{Y}^2\left[\frac{1}{n}(C_{02}-\alpha^2C_{20}) + \frac{1}{n^2}\{2(C_{02}-\alpha^2C_{20}) + \frac{\alpha^2C_{20}}{4}[C_{20}(7\alpha^2+6\alpha-21) + 20C_{02}]\right.$$

$$+ \alpha[9\alpha(1+\alpha)+2]C_{21} - [\alpha^2(3+4\alpha)C_{30} + (1+2\alpha)C_{12}]$$

$$- \frac{1}{C_{20}}[\alpha(\alpha C_{40} + 2C_{31}) + C_{22}]$$

$$+ \frac{2}{C_{20}^2}[(C_{21} + \alpha C_{30})(3\alpha C_{30} + 2C_{21})$$

$$+ C_{30}(\alpha C_{30} + C_{12})]\}] \qquad (4.18)$$

We note again that to $O(\frac{1}{n})$,

$$MSE_a(\bar{y}_{2\alpha}) = \frac{\sigma_y^2}{n}(1-\rho^2) \qquad (4.19)$$

which, as stated before, is the MSE of the ordinary regression estimator, given to the same degree of approximation.

We now proceed to show the similarity between Srivastava's estimator and the ordinary regression estimator. From (4.7a), we obtain

$$\ell n\, \bar{y}_\alpha - \ell n\, \bar{Y} = \ell n(1 + \delta\bar{y}_n) + \hat{\alpha}\, \ell n(1 + \delta\bar{x}_n) \qquad (4.20)$$

Since we have assumed $|\delta\bar{y}_n| < 1$,

$$\ell n(1 + \delta\bar{y}_n) = \delta\bar{y}_n - \frac{\delta^2\bar{y}_n}{2} + \frac{\delta^3\bar{y}_n}{3} - \ldots \tag{4.21a}$$

Hence as a first approximation

$$\ell n(1 + \delta\bar{y}_n) \doteq \delta\bar{y}_n \tag{4.21b}$$

Again, we note that

$$\frac{\bar{y}_\alpha}{\bar{Y}} = \frac{\bar{y}_\alpha - \bar{Y}}{\bar{Y}} + 1 \tag{4.21c}$$

and assume that

$$\left|\frac{\bar{y}_\alpha - \bar{Y}}{\bar{Y}}\right| < 1 \ .$$

Using (4.21a) - (4.21c) in (4.20), we obtain:

(1) a first approximation of $\bar{y}_\alpha$ to $O_p(\frac{1}{\sqrt{n}})$, where O_p denotes order in probability. This approximation is

$$\hat{\bar{y}}_{\alpha} \doteq \bar{y}_n + R\,\alpha(\bar{x}_n - \bar{X})$$

$$= \bar{y}_n + \beta(\bar{X} - \bar{x}_n) \qquad (4.22)$$

which is the usual difference estimator as defined in (1.25). We could have obtained the same result from (4.8) by retaining terms of $O_p(\frac{1}{\sqrt{n}})$. (2) A second approximation obtained directly by writing down the result of our substitution of (4.21a) - (4.21c) in (4.20). This second approximation is

$$\tilde{y}_{\alpha} \doteq \bar{y}_n + R\,\hat{\alpha}(\bar{x}_n - \bar{X}) \qquad (4.23)$$

Using (4.15a), we obtain, to $O(\frac{1}{n})$

$$E(\tilde{y}_{\alpha}) = \bar{Y} + \frac{1}{n}\{\alpha(1+\alpha)C_{20} - \frac{1}{C_{20}}(C_{21} + \alpha C_{30})\} \qquad (4.24)$$

Hence, the bias of $\tilde{y}_{\alpha}$ to $O(\frac{1}{n})$ simplifies to

$$\beta(\tilde{y}_\alpha) = -\frac{\beta\bar{X}}{n}[(C_{20} - C_{11}) + (\frac{C_{21}}{C_{11}} - \frac{C_{30}}{C_{20}})] \qquad (4.25)$$

which apart from a factor $\frac{1}{2}$, omitted from the first term on the r.h.s. is the same as (4.17b). Alternatively, we note that to $O(\frac{1}{n})$,

$$E(\bar{y}_\alpha) = \frac{1}{2}\{E(\tilde{y}_\alpha) + E(\bar{y}_R)\} \quad . \qquad (4.26)$$

If $C_{20} = C_{11}$ (i.e. $\rho = \frac{C_x}{C_y}$) and the approximations made above are valid, then $\bar{y}_\alpha$, $\tilde{y}_\alpha$ and $\bar{y}_R$ (defined in Chapter I) have the same bias.

V. NUMERICAL COMPARISONS

A. Source and Scope of Data

To investigate further the properties of the ratio and regression type estimators considered in Chapters II and III, numerical illustrations were considered. These were based on data from the Pacific Northwest on sawtimber trees of mixed species (mostly Douglas-fir), provided by George B. Hartman of the Bureau of Land Management, Portland, Oregon, through Professor Kenneth Ware of the Forestry Department at Iowa State University. In addition, Monte Carlo (simulation) techniques were applied to the same data. The data consist of three groups of trees, each group corresponding to a geographic area. The sizes of the groups are - Group 1: 1,112; Group 2: 1,094; Group 3: 958. For convenience of notation, we shall refer to the totality of trees in all three groups as Group 0. Thus Group 0 comprises 3,164 trees.

One dependent variable, Y, and five independent variables, X_i, $i = 1, 2, \ldots, 5$ defined below were considered:

Y = gross volume

X_1 = diameter breast-height, recorded by 4 inch classes

X_2 = height

$X_3 = X_1^2$

$X_4 = X_1^2 X_2$

$X_5 = H^*$, where $H^* = \begin{cases} 3 & \text{for } 0 < X_2 \leq 5 \\ 7 & \text{for } 5 < X_2 \leq 10 \end{cases}$.

All X_1's and X_2's were rounded off to integers in the field. Five different populations (indexed by 1, 2, ..., 5) were defined by each of the combinations (X_i, Y), $i = 1, 2, \ldots, 5$. The jth group of the ith population is denoted by π_{ij}, where $i = 1, 2, \ldots, 5$ and $j = 0, 1, 2, 3$. Owing to marked similarities among groups belonging to the same population, we shall consider only the population-group combinations π_{i0}, $i = 1, 2, \ldots, 5$. For brevity, we shall refer to them simply as Population 1, 2 etc.

B. Characteristics of the Populations

As a guide to the interpretation of the numerical results discussed in this chapter, we provide the following information about the five populations, Table 5.1 gives the values of the population parameters of interest. The scatter diagrams in Figures 5.1(a), ..., 5.1(d) indicate the distribution of (X_i, Y), $i = 1, 2, \ldots, 4$. We note that (X_5, Y) defines a pair of parallel lines. In Tables 5.2(a), ..., 5.2(d), the marginal frequency distribution of X_i, $i = 1, 2, \ldots, 5$ are given. We note that X_1, X_3 and X_4 are highly skewed.

We note from Table 5.1 that for these populations $C_{20} < C_{11}$. However, the converse was assumed for most of our theoretical discussion in Chapter II; thus the conclusions reached here may differ somewhat from those outlined in Chapter II.

C. Comparison of biases

Before we compare the biases and MSE's of the various estimators, we list them again for ease of reference. We note that for each

Table 5.1. Parameter Values

Parameter	Population 1	Population 2	Population 3	Population 4	Population 5
$\overline{X}$	0.22095×10^{2}	0.38764×10^{1}	0.61535×10^{3}	0.34765×10^{4}	0.41075×10^{1}
$\overline{Y}$	0.89633×10^{1}	0.89633×10^{1}	0.89633×10^{1}	0.89633×10^{1}	0.89633×10^{1}
$\overline{R}$	0.27694×10^{0}	0.16668×10^{1}	0.10476×10^{-1}	0.32704×10^{-2}	0.16526×10^{1}
R	0.40566×10^{0}	0.23123×10^{1}	0.14566×10^{-1}	0.25782×10^{-2}	0.21822×10^{1}
σ_{rY}	0.32976×10^{1}	0.21709×10^{2}	0.50291×10^{-1}	-0.55704×10^{-2}	0.23392×10^{2}
σ_{rX}	0.28433×10^{1}	0.25011×10^{1}	0.25154×10^{1}	-0.24080×10^{1}	0.21743×10^{1}
σ_{r}^{2}	0.67497×10^{-1}	0.28687×10^{1}	0.20493×10^{-4}	0.28755×10^{-5}	0.35069×10^{1}
σ_{X}^{2}	0.12716×10^{3}	0.43731×10^{1}	0.44419×10^{6}	0.25115×10^{8}	0.32019×10^{1}
σ_{XY}	0.13895×10^{3}	0.21036×10^{2}	0.85476×10^{4}	0.65303×10^{5}	0.18110×10^{2}
σ_{Y}^{2}	0.17069×10^{3}	0.17069×10^{3}	0.17069×10^{3}	0.17069×10^{3}	0.17069×10^{3}
ρ	0.94314×10^{0}	0.76995×10^{0}	0.98164×10^{0}	0.99738×10^{0}	0.77466×10^{0}
C_{20}	0.26046×10^{0}	0.29103×10^{0}	0.11731×10^{1}	0.20780×10^{1}	0.18979×10^{0}
C_{11}	0.70158×10^{0}	0.60543×10^{0}	0.15497×10^{1}	0.20956×10^{1}	0.49190×10^{0}
C_{02}	0.21246×10^{1}	0.21246×10^{1}	0.21246×10^{1}	0.21246×10^{1}	0.21246×10^{1}
C_{30}	0.15922×10^{0}	0.38408×10^{-1}	0.26150×10^{1}	0.70630×10^{1}	0.82527×10^{-1}

Table 5.1 (Continued).

Parameter	Population 1	Population 2	Population 3	Population 4	Population 5
C_{21}	0.55062×10^{0}	0.24122×10^{0}	0.36282×10^{1}	0.72295×10^{1}	0.21374×10^{0}
C_{12}	0.20087×10^{1}	0.13500×10^{1}	0.52014×10^{1}	0.74273×10^{1}	0.10676×10^{1}
C_{03}	0.76586×10^{1}	0.76586×10^{1}	0.76586×10^{1}	0.76586×10^{1}	0.76586×10^{1}
C_{40}	0.25287×10^{0}	0.16512×10^{0}	0.10754×10^{2}	0.43775×10^{2}	0.71932×10^{-1}
C_{31}	0.85050×10^{0}	0.41277×10^{0}	0.15133×10^{2}	0.45013×10^{2}	0.18629×10^{0}
C_{22}	0.30929×10^{1}	0.15635×10^{1}	0.21934×10^{2}	0.46408×10^{2}	0.86715×10^{0}
C_{13}	0.12028×10^{2}	0.79411×10^{1}	0.32649×10^{2}	0.47985×10^{2}	0.56509×10^{1}
C_{04}	0.49782×10^{2}	0.49782×10^{2}	0.49782×10^{2}	0.49782×10^{2}	0.49782×10^{2}
C_{50}	0.31945×10^{0}	0.59458×10^{-1}	0.46937×10^{2}	0.31398×10^{3}	0.46918×10^{-1}
C_{41}	0.11768×10^{1}	0.28246×10^{0}	0.67912×10^{2}	0.32472×10^{3}	0.12156×10^{0}
C_{32}	0.46105×10^{1}	0.14181×10^{1}	0.10095×10^{3}	0.33650×10^{3}	0.57964×10^{0}
C_{23}	0.19077×10^{2}	0.80062×10^{1}	0.15374×10^{3}	0.34946×10^{3}	0.39110×10^{1}
C_{14}	0.83191×10^{2}	0.52003×10^{2}	0.23927×10^{3}	0.36392×10^{3}	0.34426×10^{2}
C_{05}	0.38026×10^{3}	0.38026×10^{3}	0.38026×10^{3}	0.38026×10^{3}	0.38026×10^{3}

Table 5.1 (Continued).

Parameter	Population 1	Population 2	Population 3	Population 4	Population 5
β	0.10927×10^{1}	0.48103×10^{1}	0.19243×10^{-1}	0.26002×10^{-2}	0.56561×10^{1}
T	0.28561×10^{1}	0.27019×10^{1}	0.13458×10^{1}	0.10112×10^{1}	0.33458×10^{1}

Figure 5. 1(a). Scatter Diagram for Population-Group Combination π_{11}

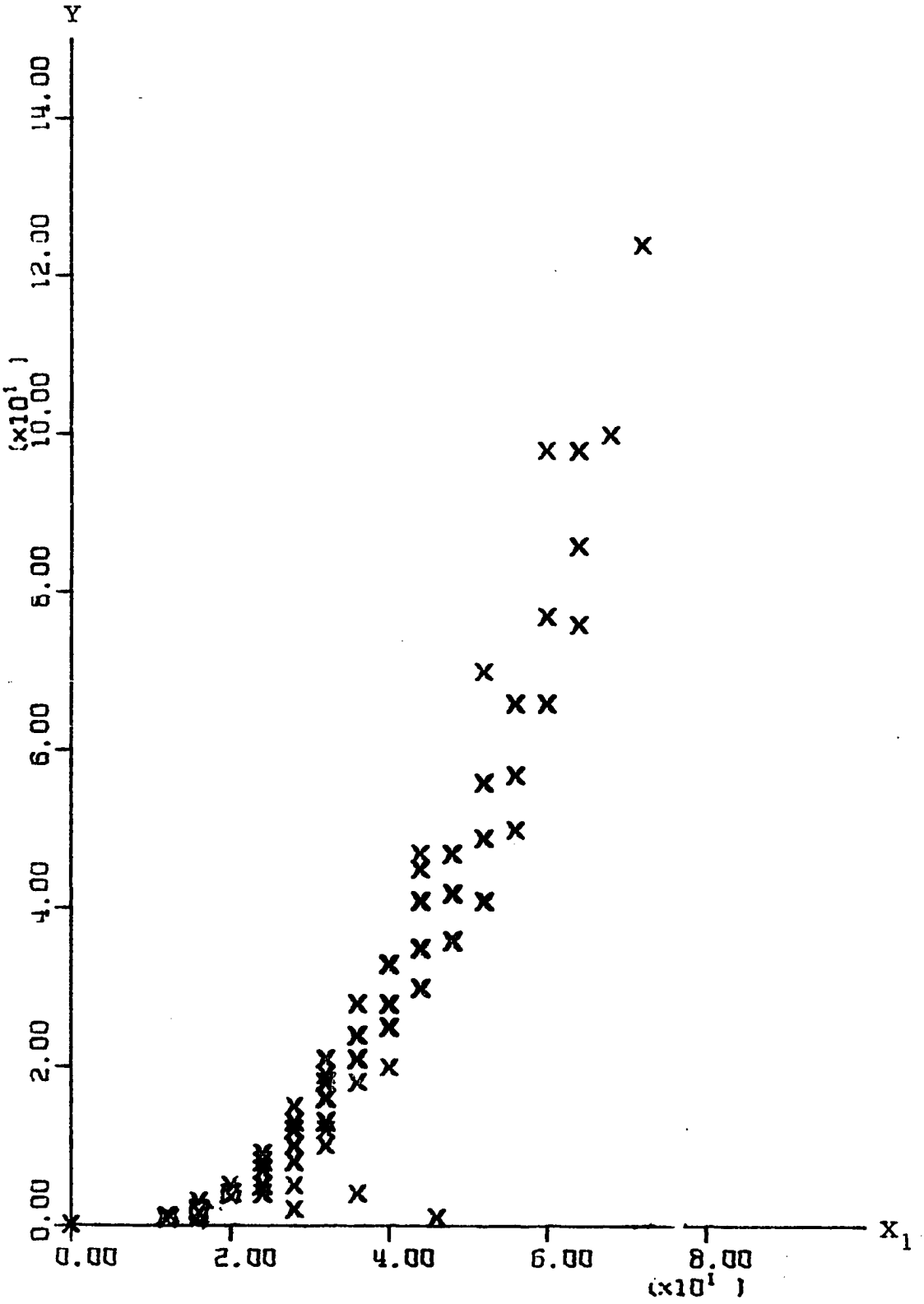
Y
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
(x10¹)
0.00
2.00
4.00
6.00
8.00
(x10¹)
X1

Figure 5.1(b). Scatter Diagram for Population-Group Combination π_{21}

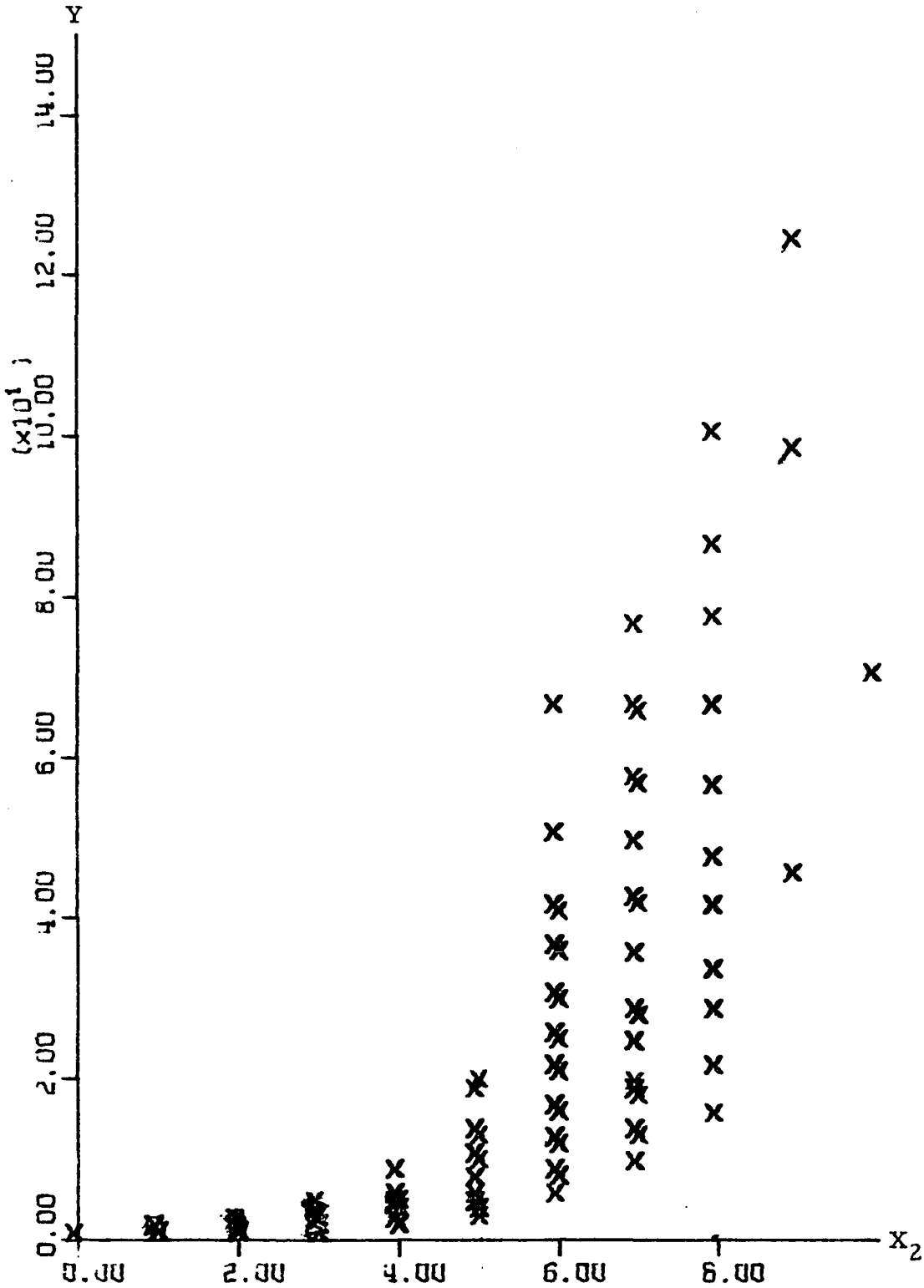
Y
(x10^1)
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
0.00
2.00
4.00
6.00
8.00
X2

Figure 5.1(c). Scatter Diagram for Population-Group Combination π_{31}

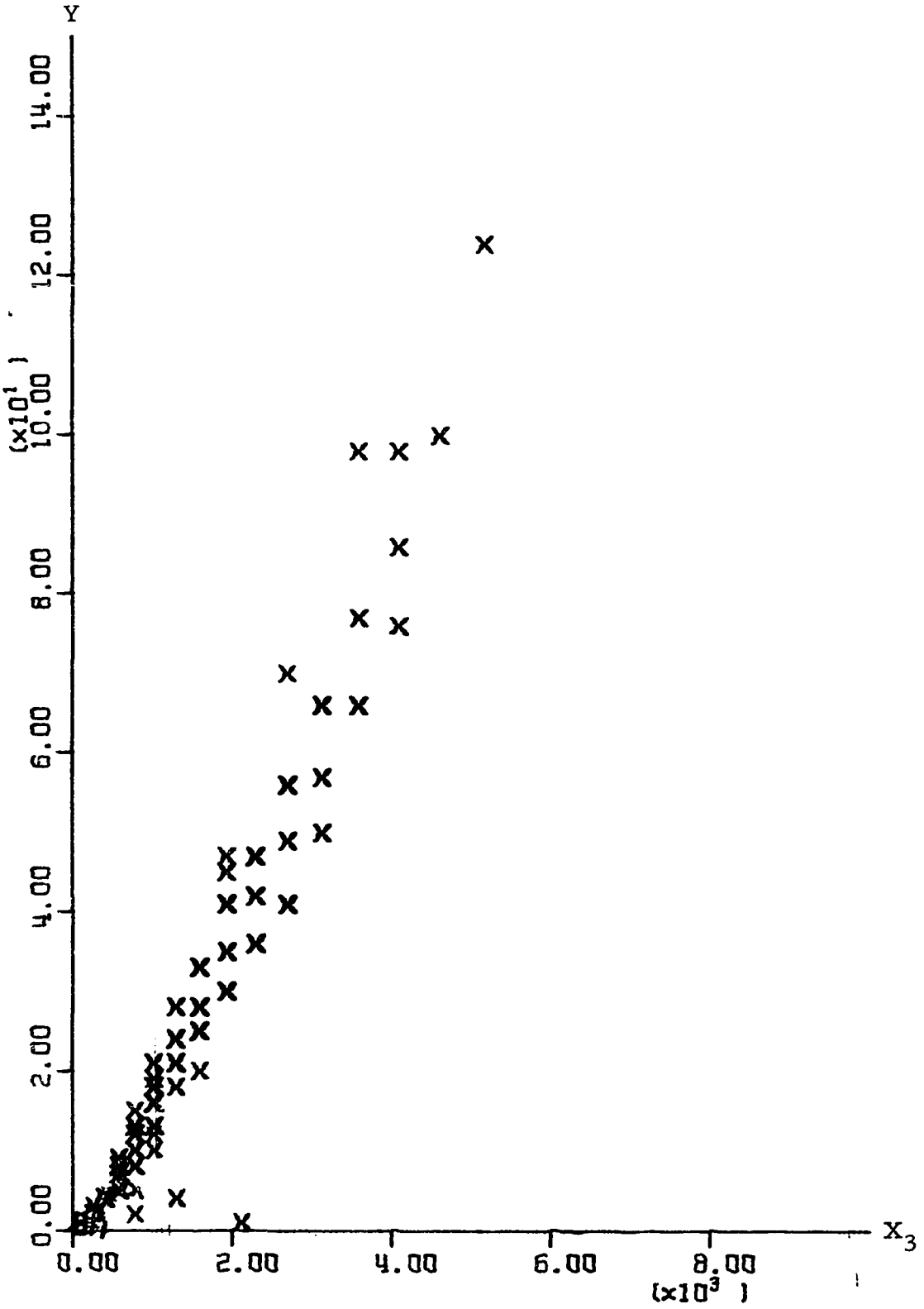
Y
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
(x10¹)
0.00
2.00
4.00
6.00
8.00
(x10³)
X3

Figure 5.1(d). Scatter Diagram for Population-Group Combination π_{41}

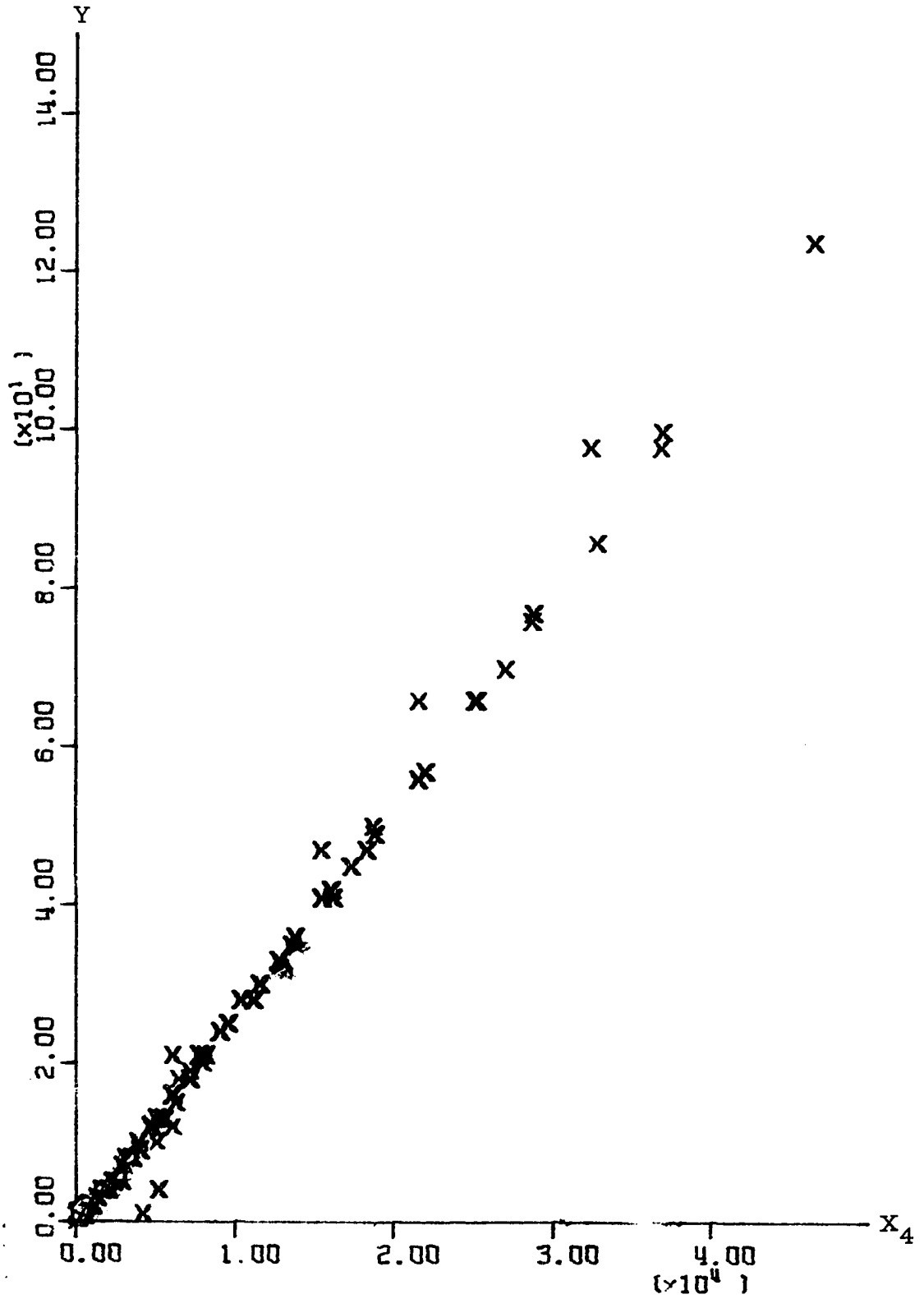
Y
14.00
12.00
10.00
8.00
6.00
4.00
2.00
0.00
(x10¹)
0.00
1.00
2.00
3.00
4.00
(x10⁴)
X4

Table 5.2(a). Marginal Frequency Distribution of X_1

Value	Frequency	Value	Frequency	Value	Frequency
12	976	36	136	60	10
16	640	40	140	64	5
20	369	44	106	68	1
24	248	48	68	72	1
28	225	52	31	Total	3, 164
32	184	56	24		

Table 5.2(b). Marginal Frequency Distribution of X_2

Value	Frequency	Value	Frequency	Value	Frequency
1	503	5	318	9	7
2	532	6	470	10	1
3	427	7	282	Total	3, 164
4	508	8	116		

Table 5.2(c). Marginal Frequency Distribution of X_5

Value	Frequency
3	2,288
7	876
Total	3,164

Table 5.2(d). Marginal Frequency Distribution of X_4

Class	Frequency	Class	Frequency	Class	Frequency
1- 999	1420	7000- 7999	106	14000-14999	-
1000-1999	467	8000- 8999	11	15000-15999	31
2000-2999	279	9000- 9999	123	16000-16999	29
3000-3999	138	10000-10999	9	17000-17999	3
4000-4999	122	11000-11999	87	18000-18999	35
5000-5999	68	12000-12999	23	19000-19999	-
6000-6999	104	13000-13999	60	20000+	49
				Total	3,164

estimator $\bar{y}_{ij}$, i = 1, 2, i = 1 refers to Scheme I (i.e. independent subsampling) while i = 2 refers to Scheme II (i.e. dependent subsampling). The estimators are

(1) $\bar{y}_{11}$, $\bar{y}_{21}$: the classical ratio estimator;

(2) $\bar{y}_{12}$, $\bar{y}_{22}$: the Hartley-Ross estimator;

(3) $\bar{y}_{13}$, $\bar{y}_{23A}$, $\bar{y}_{23B}$: the Pascual estimator;

(4) $\bar{y}_{14}$, $\bar{y}_{24}$: Beale's estimator;

(5) $\bar{y}_{15}$, $\bar{y}_{25}$: Tin's estimator;

(6) $\bar{y}_{16}$, $\bar{y}_{26}$: Quenouille's estimator;

(7) $\bar{y}_{1R}$, $\bar{y}_{2R}$: the linear regression estimator.

The comparisons in this section and Section D relate to the asymptotic expansions considered in Chapters II and III. In Tables 5.3(a), ..., 5.3(e), we give the numerical coefficients of the fractions $\frac{1}{n}$, $\frac{1}{n'}$, $\frac{1}{n^2}$, $\frac{1}{nn'}$, $\frac{1}{(n')^2}$ in the asymptotic expansions for the biases of the estimators discussed in Chapters II and III. The population-group combinations considered are, as previously stated, π_{i0} (i = 1, 2, ..., 5). These numerical coefficients are given in these tables (and in Tables 5.6(a), ..., 5.6(e)) to make it possible for subsequent comparisons, involving values of (n, n') different from those considered in this chapter, to be made.

We considered the average ranking of the biases of these estimators for specified values of n and n'. These were: (n, n') = (8, 80), (10, 30), (20, 80), (25, 75), (30, 50) and (100, 300). We note that except for the last value of n we have used an ultimate sampling fraction (n/N) of less than 0.01. Such sampling fractions are common in forest

Table 5.3(a). Coefficients of Terms in Asymptotic Expansions of Biases of Estimators-Population 1

Term / Estimator	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	-3.9540	-	0.4188	-	-
$\bar{y}_{12}$	-	-	-	-	-
$\bar{y}_{13}$	-1.1097	-	-0.6909	-	-
$\bar{y}_{14}$	-	-	-4.9570	-	-
$\bar{y}_{15}$	-	-	-3.9272	-	-
$\bar{y}_{16}$	-	-	-0.8377	-	-
$\bar{y}_{1R}$	-4.1898	-	-13.7547	-	-
$\bar{y}_{21}$	-3.9540	3.9540	0.4188	-0.4188	-
$\bar{y}_{22}$	-	-	-	-	-
$\bar{y}_{23A}$	-1.1097	3.9540	-0.6909	0.6909	-
$\bar{y}_{23B}$	-1.1097	1.1097	-0.6909	1.8006	-1.1097
$\bar{y}_{24}$	-	-	-4.9570	6.4056	-1.4486
$\bar{y}_{25}$	-	-	-3.9272	4.3460	-0.4188
$\bar{y}_{26}$	-	-	-0.8377	0.8377	-
$\bar{y}_{2R}$	-4.1898	4.1898	-13.7547	13.7547	-

Table 5.3(b). Coefficients of Terms in Asymptotic Expansions of Biases of Estimators-Population 2

Term / Estimator	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	-2.8181	-	-0.6425	-	-
$\bar{y}_{12}$	-	-	-	-	-
$\bar{y}_{13}$	-0.3159	-	-0.9585	-	-
$\bar{y}_{14}$	-	-	-1.9955	-	-
$\bar{y}_{15}$	-	-	-1.1754	-	-
$\bar{y}_{16}$	-	-	1.2851	-	-
$\bar{y}_{1R}$	-4.9686	-	-9.7245	-	-
$\bar{y}_{21}$	-2.8181	2.8181	-0.6425	0.6425	-
$\bar{y}_{22}$	-	-	-	-	-
$\bar{y}_{23A}$	-0.3159	2.8181	-0.9585	0.9585	-
$\bar{y}_{23B}$	-0.3159	0.3159	-0.9585	1.2744	-0.3159
$\bar{y}_{24}$	-	-	-1.9955	2.1731	-0.1776
$\bar{y}_{25}$	-	-	-1.1754	0.5328	0.6425
$\bar{y}_{26}$	-	-	1.2851	-1.2851	-
$\bar{y}_{2R}$	-4.9686	4.9686	-9.7245	9.7245	-

Table 5.3(c). Coefficients of Terms in Asymptotic Expansions of Biases of Estimators-Population 3

Estimator \ Term	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	-3.3759	-	-2.7982	-	-
$\bar{y}_{12}$	-	-	-	-	-
$\bar{y}_{13}$	-0.8593	-	-3.6575	-	-
$\bar{y}_{14}$	-	-	-10.2442	-	-
$\bar{y}_{15}$	-	-	-6.2841	-	-
$\bar{y}_{16}$	-	-	5.5965	-	-
$\bar{y}_{1R}$	-1.3271	-	-2.7619	-	-
$\bar{y}_{21}$	-3.3759	3.3759	-2.7982	2.7982	-
$\bar{y}_{22}$	-	-	-	-	-
$\bar{y}_{23A}$	-0.8593	3.3759	-3.6575	3.6575	-
$\bar{y}_{23B}$	-0.8593	0.8593	-3.6575	4.5168	-0.8593
$\bar{y}_{24}$	-	-	-10.2442	11.4061	-1.1619
$\bar{y}_{25}$	-	-	-6.2841	3.4859	2.7982
$\bar{y}_{26}$	-	-	5.5965	-5.5965	-
$\bar{y}_{2R}$	-1.3271	1.3271	-2.7619	2.7619	-

Table 5.3(d). Coefficients of Terms in Asymptotic Expansions of Biases of Estimators-Population 4

Term / Estimator	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	-0.1586	-	0.5034	-	-
$\bar{y}_{12}$	-	-	-	-	-
$\bar{y}_{13}$	-2.5649	-	-2.0616	-	-
$\bar{y}_{14}$	-	-	-2.3247	-	-
$\bar{y}_{15}$	-	-	-1.9952	-	-
$\bar{y}_{16}$	-	-	-1.0067	-	-
$\bar{y}_{1R}$	-0.4586	-	-1.4839	-	-
$\bar{y}_{21}$	-0.1586	0.1586	0.5034	-0.5034	-
$\bar{y}_{22}$	-	-	-	-	-
$\bar{y}_{23A}$	-2.5666	0.1586	-2.0616	2.0616	-
$\bar{y}_{23B}$	-2.5666	2.5666	-2.0616	4.6265	-2.5649
$\bar{y}_{24}$	-	-	-2.3247	3.1575	-0.8329
$\bar{y}_{25}$	-	-	-1.9952	2.4986	-0.5034
$\bar{y}_{26}$	-	-	-1.0067	1.0067	-
$\bar{y}_{2R}$	-0.4586	0.4586	-1.4835	1.4835	-

Table 5.3(e). Coefficients of Terms in Asymptotic Expansions of Biases of Estimators-Population 5

Estimator \ Term	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	-2.7080	-	-0.3657	-	-
$\bar{y}_{12}$	-	-	-	-	-
$\bar{y}_{13}$	-0.5325	-	-0.8982	-	-
$\bar{y}_{14}$	-	-	-1.3243	-	-
$\bar{y}_{15}$	-	-	-0.8104	-	-
$\bar{y}_{16}$	-	-	0.7315	-	-
$\bar{y}_{1R}$	0.0077	0.0000	0.0594	-	-
$\bar{y}_{21}$	-2.7080	2.7080	-0.3657	0.3657	-
$\bar{y}_{22}$	-	-	-	-	-
$\bar{y}_{23A}$	-0.5325	2.7080	-0.8982	0.8982	-
$\bar{y}_{23B}$	-0.5325	0.5325	-0.8982	1.4307	-0.5325
$\bar{y}_{24}$	-	-	-1.3243	1.4725	-0.1482
$\bar{y}_{25}$	-	-	-0.8104	0.4446	0.3657
$\bar{y}_{26}$	-	-	0.7315	-0.7315	-
$\bar{y}_{2R}$	0.0077	-0.0077	0.0594	-0.0594	-

surveys. Also for n large, the biases of the ratio and regression estimators are likely to be unimportant and therefore the comparisons uninteresting. The average ranking of the estimators (from lowest to highest in absolute value of bias) is given in Table 5.4 separately for each scheme. Two estimators shown in brackets have the same average rank.

More general conclusions can be deduced from Tables 5.3(a), ..., 5.3(e). We recall that $\bar{y}_{12}$ and $\bar{y}_{22}$ are unbiased. We also note that for Scheme I, for $n \gtrless \alpha$ (a constant depending on the population-group combination being considered), we can rank the estimators in a definite order. The various orders for Scheme I are as follows:

(1) For Populations 1 and 2 (with $n \geq 4$), for Population 3 (with $n \geq 8$) and for Population 5 (with $n \leq 87$) the order is the same as in Table 5.4.

(2) For Population 4 with $n \geq 18$, the order differs slightly from that shown in Table 5.4 with $\bar{y}_{11}$ being placed between $\bar{y}_{14}$ and $\bar{y}_{1R}$.

Thus for Scheme I, we conclude that for $n \geq 18$ for all populations (except Population 5), the general ranking of the estimators (from lowest to highest in absolute value of bias) is (with two modifications):

$$\bar{y}_{12}, \bar{y}_{16}, \bar{y}_{15}, \bar{y}_{14}, \bar{y}_{13}, \bar{y}_{11}, \bar{y}_{1R} .$$

The two changes occur in Population 2 where $\bar{y}_{15}$ has a smaller bias than $\bar{y}_{16}$ and in Population 4, where $\bar{y}_{13}$ has the largest. Some indication of the magnitudes of the biases is given in Table 5.5, where

Table 5.4. Average Ranking of Estimators with Respect to Bias for Selected Values of (n, n').

Population	Scheme	Ranking
1	I	$\bar{y}_{12}, \bar{y}_{16}, \bar{y}_{15}, \bar{y}_{14}, \bar{y}_{13}, \bar{y}_{11}, \bar{y}_{1R}$
	II	$\bar{y}_{22}, \bar{y}_{26}, \bar{y}_{25}, \bar{y}_{24}, \bar{y}_{23A}, \bar{y}_{23B}, \bar{y}_{21}, \bar{y}_{2R}$
2	I	$\bar{y}_{12}, \bar{y}_{15}, \bar{y}_{16}, \bar{y}_{14}, \bar{y}_{13}, \bar{y}_{11}, \bar{y}_{1R}$
	II	$\bar{y}_{22}, \bar{y}_{26}, \bar{y}_{25}, \bar{y}_{24}, \bar{y}_{23B}, \bar{y}_{23A}, \bar{y}_{21}, \bar{y}_{2R}$
3	I	$\bar{y}_{12}, \bar{y}_{16}, \bar{y}_{15}, \bar{y}_{14}, \bar{y}_{13}, \bar{y}_{1R}, \bar{y}_{11}$
	II	$\bar{y}_{22}, \bar{y}_{26}, \bar{y}_{25}, \bar{y}_{23A}, \bar{y}_{24}, \bar{y}_{23B}, \bar{y}_{2R}, \bar{y}_{21}$
4	I	$\bar{y}_{12}, \bar{y}_{16}, \bar{y}_{15}, \bar{y}_{11}, \bar{y}_{14}, \bar{y}_{1R}, \bar{y}_{13}$
	II	$\bar{y}_{22}, \bar{y}_{26}, \bar{y}_{25}, \bar{y}_{21}, \bar{y}_{24}, \bar{y}_{2R}, \bar{y}_{23B}, \bar{y}_{23A}$
5	I	$\bar{y}_{12}, \bar{y}_{1R}, \bar{y}_{16}, \bar{y}_{15}, \bar{y}_{14}, \bar{y}_{13}, \bar{y}_{11}$
	II	$\bar{y}_{22}, \bar{y}_{2R}, \bar{y}_{26}, \bar{y}_{25}, \bar{y}_{24}, \bar{y}_{23A}, \bar{y}_{23B}, \bar{y}_{21}$

Table 5.5. Bias of Estimators for (n, n') = (8, 80)

Estimator	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
$\bar{y}_{11}$	-0.487703	-0.362306	-0.465713	-0.011956	-0.344214
$\bar{y}_{12}$	-	-	-	-	-
$\bar{y}_{13}$	-0.149505	-0.054466	-0.164558	-0.352827	-0.080593
$\bar{y}_{14}$	-0.077453	-0.031180	-0.160066	-0.036324	-0.020692
$\bar{y}_{15}$	-0.061362	-0.018365	-0.098189	-0.031176	-0.012662
$\bar{y}_{16}$	-0.013088	-0.020080	0.087445	-0.015730	0.011429
$\bar{y}_{1R}$	-0.738647	-0.773016	-0.209040	-0.080506	0.001894
$\bar{y}_{21}$	-0.438932	-0.326075	-0.419142	-0.010760	-0.309792
$\bar{y}_{22}$	-	-	-	-	-
$\bar{y}_{23A}$	-0.099142	-0.017900	-0.116813	-0.347855	-0.045505
$\bar{y}_{23B}$	-0.133121	-0.048717	-0.147046	-0.314145	-0.071934
$\bar{y}_{24}$	-0.067670	-0.027812	-0.142426	-0.031520	-0.018414
$\bar{y}_{25}$	-0.054637	-0.017432	-0.092305	-0.027350	-0.011910
$\bar{y}_{26}$	-0.011780	-0.018072	0.078700	-0.014157	-0.010286
$\bar{y}_{2R}$	-0.664783	-0.695715	-0.188139	-0.072450	-0.001705

the case $(n, n') = (8, 80)$ is considered. For larger values of n, the biases are generally smaller than shown in Table 5.5.

With Scheme II, as with Scheme I, we can also draw some general conclusions with regard to ranking of the absolute values of the bias of some of the estimators. We note that for estimators whose biases are at most of $O(\frac{1}{n^2})$, the following general pattern is discernible. For all five populations, with $n' > n > 1$ (and for Population 2, $n' < 5.8n$),

$$\beta(\bar{y}_{22}) < |\beta(\bar{y}_{26})| < |\beta(\bar{y}_{25})| < |\beta(\bar{y}_{24})| \quad .$$

In conclusion, we note from Table 5.4 and from our preceding discussions that for Scheme I, $\bar{y}_{16}$ has consistently a low ranking in bias, whereas with one exception in each case, $\bar{y}_{11}$ and $\bar{y}_{1R}$ consistently obtained high ranks. There is on the average little difference (in magnitude) among $\bar{y}_{13}$, $\bar{y}_{14}$ and $\bar{y}_{15}$. The same general pattern is discernible for the corresponding estimators under Scheme II.

We note the consistency (within populations) of ranking between Scheme I and II, and consistency of general results (for Scheme I) with specific results (for Scheme I).

D. Comparison of MSE's

Following the same procedure used for ranking the biases of the estimators, we again rank the estimators with respect to M.S.E. Tables 5.6(a), ..., 5.6(e) and Table 5.7 are the MSE analogues of Tables 5.2(a), ..., 5.2(e) and Table 5.4 respectively. An equality

between two estimators in Table 5.7 implies that they have identical values in the asymptotic expansion.

We recall from Chapter II that for Scheme I, $\bar{y}_{ij}$ (j = 1, 3, 4, 5, 6) have the same MSE to $O(\frac{1}{n'})$. Hence to compare their MSE's we need consider only terms of $O(\frac{1}{n^2})$ and higher. From the examination of these terms, we conclude that for Populations 1, 2 and 5 (with $n' > n > 1$), for Population 3 (with $3n \leq n' \leq 16n$) and for Population 4 (with $n < n' \leq 23n$), the ranking of the estimators with respect to MSE is the same as in Table 5.7.

We recall that to $O(\frac{1}{n'})$, $MSE(\bar{y}_{ij})$ is identical for $j \neq 2, R$. If n is so large that only terms to $O(\frac{1}{n'})$ need be considered, then

$$MSE(\bar{y}_{1R}) < MSE(\bar{y}_{1j}) < MSE(\bar{y}_{12}), \qquad j \neq 2, R$$

for Populations 1, 2 and 5, if $n' > 2.9n$. For Population 4, the condition becomes $n' > 7.3n$ and for Population 3, $n' > 236n$ (a rather unrealistic condition).

Again with Scheme II, it is possible to make more general comparisons of MSE's for these estimators which have the same MSE to $O(\frac{1}{n'})$, namely $\bar{y}_{2j}$ ($j \neq 2, R$). For Population 1, the ranking is the same as in Table 5.7 for $n' > 3.1n$, except that the equality ($\bar{y}_{24} = \bar{y}_{25}, \bar{y}_{26}$) is replaced by the inequality $\bar{y}_{24} = \bar{y}_{25}, \bar{y}_{26}$. Again for Population 2 (with $n' \geq 3n$), Population 3 (with $2.6n \leq n' \leq 14n$), Population 4 (with $n' \leq 5.3n$) and Population 5 (with $1.9n \leq n' \leq 4.9n$), the exact ranking is as in Table 5.7.

Table 5.6(a). Coefficients of Terms in Asymptotic Expansions of MSE's of Estimators-Population 1

Estimator \ Term	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	78.8846	20.9253	-15.9529	2.0844	-
$\bar{y}_{12}$	103.4843	9.7521	16.6668	8.5826	-
$\bar{y}_{13}$	78.8846	20.9253	32.6623	15.3647	-
$\bar{y}_{14}$	78.8846	20.9253	36.1798	20.5459	-
$\bar{y}_{15}$	78.8846	20.9253	36.1798	20.5459	-
$\bar{y}_{16}$	78.8846	20.9253	72.3597	20.5459	-
$\bar{y}_{1R}$	18.8593	151.8324	75.7686	-174.7710	-
$\bar{y}_{21}$	78.8846	91.8070	-15.9529	-41.2846	57.2375
$\bar{y}_{22}$	103.4843	67.2073	16.6668	-24.7511	8.0842
$\bar{y}_{23A}$	78.8846	91.8070	32.6623	36.8839	57.2375
$\bar{y}_{23B}$	78.8846	91.8070	32.6623	-19.8156	-12.8467
$\bar{y}_{24}$	78.8846	91.8070	36.1798	-20.5459	+15.6339
$\bar{y}_{25}$	78.8846	91.8070	36.1798	-20.5459	-15.6339
$\bar{y}_{26}$	78.8846	91.8070	72.3597	-166.1632	93.8035
$\bar{y}_{2R}$	18.8593	151.8324	75.7686	-144.9500	69.1814

Table 5.6(b). Coefficients of Terms in Asymptotic Expansions of MSE's of Estimators-Population 2

Term / Estimator	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	96.7905	23.3814	-13.4174	13.4660	-
$\bar{y}_{12}$	112.7159	12.1492	18.8007	12.5453	-
$\bar{y}_{13}$	96.7905	23.3814	48.8518	26.5203	-
$\bar{y}_{14}$	96.7905	23.3814	36.1103	28.1685	-
$\bar{y}_{15}$	96.7905	23.3814	36.1103	28.1685	-
$\bar{y}_{16}$	96.7905	23.3814	72.2207	28.1685	-
$\bar{y}_{1R}$	69.5014	101.1902	120.2982	75.2460	-
$\bar{y}_{21}$	96.7905	73.9010	-13.4174	-5.6502	19.0676
$\bar{y}_{22}$	112.7159	57.9758	18.8007	-25.0561	6.2554
$\bar{y}_{23A}$	96.7905	73.9010	48.8518	41.2877	19.0676
$\bar{y}_{23B}$	96.7905	73.9010	48.8518	-27.2369	-21.6149
$\bar{y}_{24}$	96.7905	73.9010	36.1103	-28.1685	-7.9418
$\bar{y}_{25}$	96.7905	73.9010	36.1103	-28.1685	-7.9418
$\bar{y}_{26}$	96.7905	73.9010	72.2206	-119.8717	47.6511
$\bar{y}_{2R}$	69.5014	101.1902	120.2981	-200.3188	80.0207

Table 5.6(c). Coefficients of Terms in Asymptotic Expansions of MSE's of Estimators-Population 3

Estimator \ Term	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	15.9269	94.2455	34.4710	-52.3097	-
$\bar{y}_{12}$	40.3486	48.7522	15.4302	9.1028	-
$\bar{y}_{13}$	15.9269	94.2455	38.8660	0.6133	-
$\bar{y}_{14}$	15.9269	94.2455	30.0800	18.6832	-
$\bar{y}_{15}$	15.9269	94.2455	30.0800	18.6832	-
$\bar{y}_{16}$	15.9269	94.2455	60.1601	18.6832	-
$\bar{y}_{1R}$	6.2115	164.4802	24.7325	-118.1203	-
$\bar{y}_{21}$	15.9269	154.7647	34.4701	-78.0945	43.6235
$\bar{y}_{22}$	40.3486	130.3430	15.4302	-21.7576	6.3274
$\bar{y}_{23A}$	15.9269	154.7647	38.8660	24.2304	43.6235
$\bar{y}_{23B}$	15.9269	154.7647	38.8660	13.5054	-52.3714
$\bar{y}_{24}$	15.9269	154.7647	30.0800	-18.6832	-11.3968
$\bar{y}_{25}$	15.9269	154.7647	30.0800	-18.6832	-11.3968
$\bar{y}_{26}$	15.9269	154.7647	60.1601	-128.5415	68.3814
$\bar{y}_{2R}$	6.2115	164.4802	24.7316	-55.5765	30.8439

Table 5.6(d). Coefficients of Terms in Asymptotic Expansions of MSE's of Estimators-Population 4

Estimator \ Term	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	0.9036	166.9454	0.7466	-4.0289	-
$\bar{y}_{12}$	12.1731	268.6169	78.0155	72.2170	-
$\bar{y}_{13}$	0.9036	166.9454	4.6712	-93.6677	-
$\bar{y}_{14}$	0.9036	166.9454	1.9030	1.8776	-
$\bar{y}_{15}$	0.9036	166.9454	1.9030	1.8776	-
$\bar{y}_{16}$	0.9036	166.9454	3.8060	1.8776	-
$\bar{y}_{1R}$	0.8915	169.8001	4.8620	-62.0728	-
$\bar{y}_{21}$	0.9036	169.7880	0.7466	-15.7285	14.9819
$\bar{y}_{22}$	12.1731	158.5186	78.0155	-83.8140	5.7985
$\bar{y}_{23A}$	0.9036	169.7880	4.6712	-60.8814	14.9819
$\bar{y}_{23B}$	0.9036	169.7880	4.6712	-70.6070	65.9358
$\bar{y}_{24}$	0.9036	169.7880	1.9030	-1.8779	-0.0251
$\bar{y}_{25}$	0.9036	169.7880	1.9030	-1.8779	-0.0251
$\bar{y}_{26}$	0.9036	169.7880	3.8060	-3.9568	0.1508
$\bar{y}_{2R}$	0.8915	169.8001	4.8742	-16.3855	11.5125

Table 5.6(e). Coefficients of Terms in Asymptotic Expansions of MSE's of Estimators-Population 5

Estimator \ Term	$\frac{1}{n}$	$\frac{1}{n'}$	$\frac{1}{n^2}$	$\frac{1}{nn'}$	$\frac{1}{(n')^2}$
$\bar{y}_{11}$	106.8986	15.2477	-11.2502	11.0746	-
$\bar{y}_{12}$	119.5795	8.7443	15.9567	11.2290	-
$\bar{y}_{13}$	106.8986	15.2477	42.3090	18.4763	-
$\bar{y}_{14}$	106.8986	15.2477	27.6211	20.2878	-
$\bar{y}_{15}$	106.8986	15.2477	27.6211	20.2878	-
$\bar{y}_{16}$	106.8986	15.2477	55.2422	20.2878	-
$\bar{y}_{1R}$	68.2592	102.4324	162.9093	163.1731	-
$\bar{y}_{21}$	106.8986	63.7930	-11.2502	-6.0730	17.3233
$\bar{y}_{22}$	119.5795	51.1121	15.9567	-20.6844	4.7277
$\bar{y}_{23A}$	106.8986	63.7930	42.3090	34.4541	17.3233
$\bar{y}_{23B}$	106.8986	63.7930	42.3090	-23.8344	-18.4746
$\bar{y}_{24}$	106.8986	63.7930	27.6211	-20.2878	-7.3333
$\bar{y}_{25}$	106.8986	63.7930	27.6211	-20.2878	-7.3333
$\bar{y}_{26}$	106.8986	63.7930	55.2422	-99.2414	43.9993
$\bar{y}_{2R}$	68.2592	102.4324	162.9106	-162.8963	-0.0135

Table 5.7. Average Ranking of Estimators with Respect to MSE for Selected Values of (n, n')

Population	Scheme	Ranking
1	I	$(\bar{y}_{1R}, \bar{y}_{11}), \bar{y}_{13}, \bar{y}_{14} = \bar{y}_{15}, \bar{y}_{16}, \bar{y}_{12}$
	II	$\bar{y}_{2R}, \bar{y}_{21}, \bar{y}_{23B}, (\bar{y}_{24} = \bar{y}_{25}, \bar{y}_{26}), \bar{y}_{23A}, \bar{y}_{22}$
2	I	$\bar{y}_{11}, \bar{y}_{14} = \bar{y}_{15}, \bar{y}_{1R}, \bar{y}_{13}, \bar{y}_{16}, \bar{y}_{12}$
	II	$\bar{y}_{2R}, \bar{y}_{21}, \bar{y}_{24} = \bar{y}_{25}, \bar{y}_{23B}, \bar{y}_{26}, \bar{y}_{23A}, \bar{y}_{22}$
3	I	$\bar{y}_{11}, \bar{y}_{14} = \bar{y}_{15}, \bar{y}_{13}, \bar{y}_{16}, \bar{y}_{12}, \bar{y}_{1R}$
	II	$\bar{y}_{2R}, \bar{y}_{21}, \bar{y}_{24} = \bar{y}_{25}, \bar{y}_{26}, \bar{y}_{23B}, \bar{y}_{23A}, \bar{y}_{22}$
4	I	$\bar{y}_{13}, \bar{y}_{11}, y_{14} = \bar{y}_{15}, \bar{y}_{1R}, \bar{y}_{16}, \bar{y}_{12}$
	II	$\bar{y}_{23A}, \bar{y}_{23B}, \bar{y}_{21}, \bar{y}_{2R}, \bar{y}_{24} = \bar{y}_{25}, \bar{y}_{26}, \bar{y}_{22}$
5	I	$\bar{y}_{11}, \bar{y}_{14} = \bar{y}_{15}, \bar{y}_{1R}, \bar{y}_{13}, \bar{y}_{16}, \bar{y}_{12}$
	II	$\bar{y}_{2R}, \bar{y}_{21}, \bar{y}_{24} = \bar{y}_{25}, \bar{y}_{26}, \bar{y}_{23B}, \bar{y}_{23A}, \bar{y}_{22}$

In general, for $\bar{y}_{2j}$ ($j \neq 2, R$), $\bar{y}_{26}$ is worse than $\bar{y}_{24} = \bar{y}_{25}$; $\bar{y}_{21}$ is better than the rest but as noted in Section C, $\bar{y}_{21}$ has a relatively larger bias. Hence if one is concerned about bias, one would normally use either $\bar{y}_{24}$ or $\bar{y}_{25}$. The estimator $\bar{y}_{23B}$ is usually only slightly inferior to $\bar{y}_{21}$ and $\bar{y}_{24} = \bar{y}_{25}$ but $\bar{y}_{23A}$ is usually the worst of the six estimators we are discussing here.

We can make further general conclusions for Scheme II if we assume that n is so large that we need consider terms to $O(\frac{1}{n'})$ only. We recall from the theory in Chapter II that to $O(\frac{1}{n'})$, $MSE(\bar{y}_{2j})$ is identical

for $j \neq 2$, R. It is easy to show that to $O(\frac{1}{n'})$ for all five populations considered,

$$MSE(\bar{y}_{2R}) < MSE(\bar{y}_{2j}) < MSE(\bar{y}_{22}) \qquad j \neq 2, R \quad .$$

The comparison involving $\bar{y}_{22}$ and $\bar{y}_{2j}$ (j = 1, 3A, 3B, 4, 5, 6) supports results previously obtained in Chapter II (see 2.88c). We note that

(1) For Populations 1, 2, 3 and 5,

$$\bar{R} < R \text{ but } \rho > \frac{(\bar{R} + R)}{2} \frac{\sigma_X}{\sigma_Y} ;$$

and

(2) for Population 4,

$$\bar{R} > R \text{ but } \rho < \frac{(\bar{R} + R)}{2} \frac{\sigma_X}{\sigma_Y} \quad .$$

We note from Table 5.6 and our previous discussion in this section that for Scheme I $\bar{y}_{11}$ has consistently a low rank whereas $\bar{y}_{1R}$ is subject to considerable fluctuations in rank from one population to another. However, $\bar{y}_{12}$ consistently performs poorly. For Scheme II, $\bar{y}_{2R}$ and $\bar{y}_{21}$ are consistelyly good, $\bar{y}_{23A}$ is mostly poor and $\bar{y}_{22}$ is consistelyly poor. There is not much difference in terms of rank and magnitude of MSE among $\bar{y}_{23B}$, $\bar{y}_{24}$, $\bar{y}_{25}$, all of which perform moderately well. In general, for large values of n and n' e.g. (n, n') = (25, 75) and (100, 300), there is very little to choose among the ratio-type estimators (excluding $\bar{y}_{12}$ and $\bar{y}_{22}$) - see Tables 5.8(a) and 5.8(b).

Table 5.8(a). MSE's of Estimators for (n, n') = (25, 75)

Estimator	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
$\bar{y}_{11}$	3.40997	4.16908	1.92094	2.26113	4.46715
$\bar{y}_{12}$	4.30064	4.70739	2.29352	4.23182	4.93129
$\bar{y}_{13}$	3.49484	4.27568	1.95619	2.21960	4.55679
$\bar{y}_{14}$	3.50323	4.25617	1.95177	2.26613	4.53426
$\bar{y}_{15}$	3.50323	4.25617	1.95177	2.26613	4.53426
$\bar{y}_{16}$	3.56112	4.31395	1.99990	2.26917	4.57845
$\bar{y}_{1R}$	2.80682	4.36187	2.41810	2.27433	4.44381
$\bar{y}_{21}$	4.34210	4.83587	2.72186	2.29545	5.10835
$\bar{y}_{22}$	5.05037	5.29947	3.36606	2.68166	5.48001
$\bar{y}_{23A}$	4.46158	4.96054	2.78347	2.27766	5.21567
$\bar{y}_{23B}$	4.41888	4.91676	2.76068	2.28153	5.17821
$\bar{y}_{24}$	4.42362	4.89831	2.73674	2.30202	5.15858
$\bar{y}_{25}$	4.42362	4.89831	2.73674	2.30202	5.15858
$\bar{y}_{26}$	4.42330	4.91706	2.74046	2.30399	5.16979
$\bar{y}_{2R}$	2.83502	4.22912	2.45694	2.30077	4.26991

Table 5.8(b). MSE's of Estimators for (n, n') = (100, 300)

Estimator	Pop. 1	Pop. 2	Pop. 3	Pop. 4	Pop. 5
$\bar{y}_{11}$	0.857072	1.04495	0.475124	0.565461	1.11905
$\bar{y}_{12}$	1.06930	1.16995	0.567840	1.02733	1.22691
$\bar{y}_{13}$	0.862376	1.05161	0.477328	0.562865	1.12466
$\bar{y}_{14}$	0.862900	1.05039	0.477051	0.565773	1.12325
$\bar{y}_{15}$	0.862900	1.05039	0.477051	0.565773	1.12325
$\bar{y}_{16}$	0.866518	1.05400	0.480059	0.565964	1.12601
$\bar{y}_{1R}$	0.696452	1.04685	0.608918	0.573333	1.04576
$\bar{y}_{21}$	1.09253	1.21292	0.676480	0.574721	1.28049
$\bar{y}_{22}$	1.25980	1.32152	0.838851	0.655198	1.36713
$\bar{y}_{23A}$	1.10000	1.22071	0.680330	0.673601	1.28720
$\bar{y}_{23B}$	1.09733	1.21798	0.678906	0.573842	1.28486
$\bar{y}_{24}$	1.09763	1.21682	0.677410	0.575123	1.28363
$\bar{y}_{25}$	1.09763	1.21682	0.677410	0.575123	1.28363
$\bar{y}_{26}$	1.09761	1.21800	0.677642	0.575246	1.28433
$\bar{y}_{2R}$	0.698214	1.03856	0.611345	0.574985	1.03489

To summarize our results, we recall that as a working rule, the effect of bias on the accuracy of an estimator is considered to be negligible if $\beta^2/\sigma^2 < 0.01$ [9]. In the examples considered in Tables 5.8(a) and 5.8(b), except for $\bar{y}_{11}$ and $\bar{y}_{21}$ (Population 1), the bias is unimportant since $\beta^2/\sigma^2 < 0.01$ in each case. Hence for such situations if

one had to choose a ratio-type estimator, one would consistently do well by using the classical ratio estimator $(\bar{y}_{11}, \bar{y}_{21})$. This view is supported by studies done on single phase sampling involving most of these estimators, e.g. [16]. We note further that for the two exceptions mentioned above, $\beta^2/\sigma^2 < 0.04$ and even in such cases, as Cochran points out, the disturbance in the probability of error is still rather modest [9]. For small values of n where β^2/σ^2 may not satisfy the "negligibility" criterion in the case of the classical ratio estimator, one would normally prefer Tin's (or Beale's) estimator.

E. Monte Carlo Simulation

In addition to the numerical examples discussed in Sections C and D above, Monte Carlo simulation methods were applied to the estimators in Scheme II, excluding Quenouille's estimator, $\bar{y}_{26}$. A first phase sample of size n' was selected and a subsample of size n was then drawn. Three pairs of values of (n, n') were used. These were: (n, n') = (10, 30), (20, 80) and (100, 300). We shall, however, limit our discussion in this section to the case (n, n') = (10, 30). Selection at both phases was by simple random sampling (i.e., without replacement). The process was repeated 2,000 times. For reasons of economy, this Monte Carlo work was limited to Group 0 and the first four populations. Population 5 was excluded from consideration since samples having $b = s_{xy}/s_x^2$ undefined were common.

In this section, we shall compare the Monte Carlo estimates of the bias and MSE's with the corresponding results from the asymptotic expansions. Because comparisons are uninteresting when n and n'

are large, choosing $n = 10$, $n' = 30$ would seem to provide a reasonable illustrative example. To help in the interpretation of the results, we also discuss briefly the distributions of the sample values of the estimators and give some information on their skewness and kurtosis.

(1) Distribution of estimators

The frequency distributions of the 2,000 (two-phase) sample values of the estimators for Population 1 (only) are given in Table 5.9(a). In Figures 5.2(a) and 5.2(b), we give the histograms for $\bar{y}_{21}$ (Populations 1 and 3) and $\bar{y}_{2R}$ (Population 1). We selected these three for presentation because they offer, in our opinion, the best contrasts. For example, for Population 1, the histograms for $\bar{y}_{2j}$, $j = 1, 2, 3A, 3B, 4, 5$ are very similar while each differs somewhat from that of $\bar{y}_{2R}$. In Table 5.9(b), we give the estimated third and fourth moments of the distributions of the estimators together with their measures of skewness (γ_1) and kurtosis (γ_2). We recall that if a distribution is mesokurtic (i.e. shaped like a normal), then $\gamma_2 = 0$, if leptokurtic (i.e. peaked), $\gamma_2 > 0$ and if platykurtic (i.e. flat-topped), $\gamma_2 < 0$. However the converse in each case is not necessarily true [29].

We might expect that the 2,000 random estimated values of $\bar{Y}$ would, for each estimator, have a near normal distribution. The results in Table 5.9(b) suggest that while the underlying curves of the distributions are shaped approximately like normals, they are positively skewed in each case. The least skewed distributions are $\bar{y}_{24}$ (Populations 1 and 2), $\bar{y}_{25}$ (Population 3) and $\bar{y}_{2R}$ (Population 5). For all populations the distribution of $\bar{y}_{22}$ has the largest value of $\hat{\gamma}_1$.

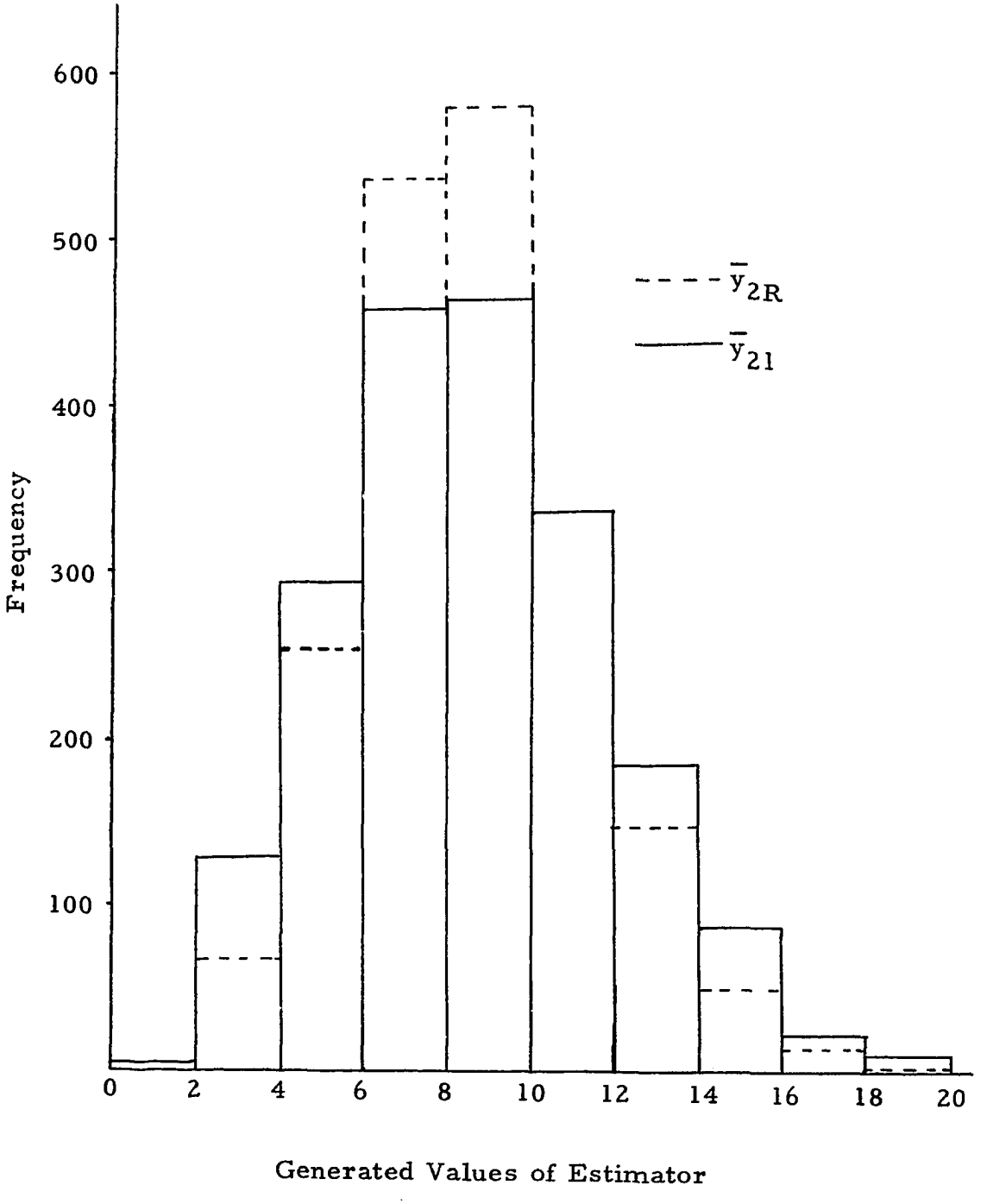

Figure 5.2(a). Histograms for $\bar{y}_{21}$ and $\bar{y}_{2R}$ (Pop. 1)

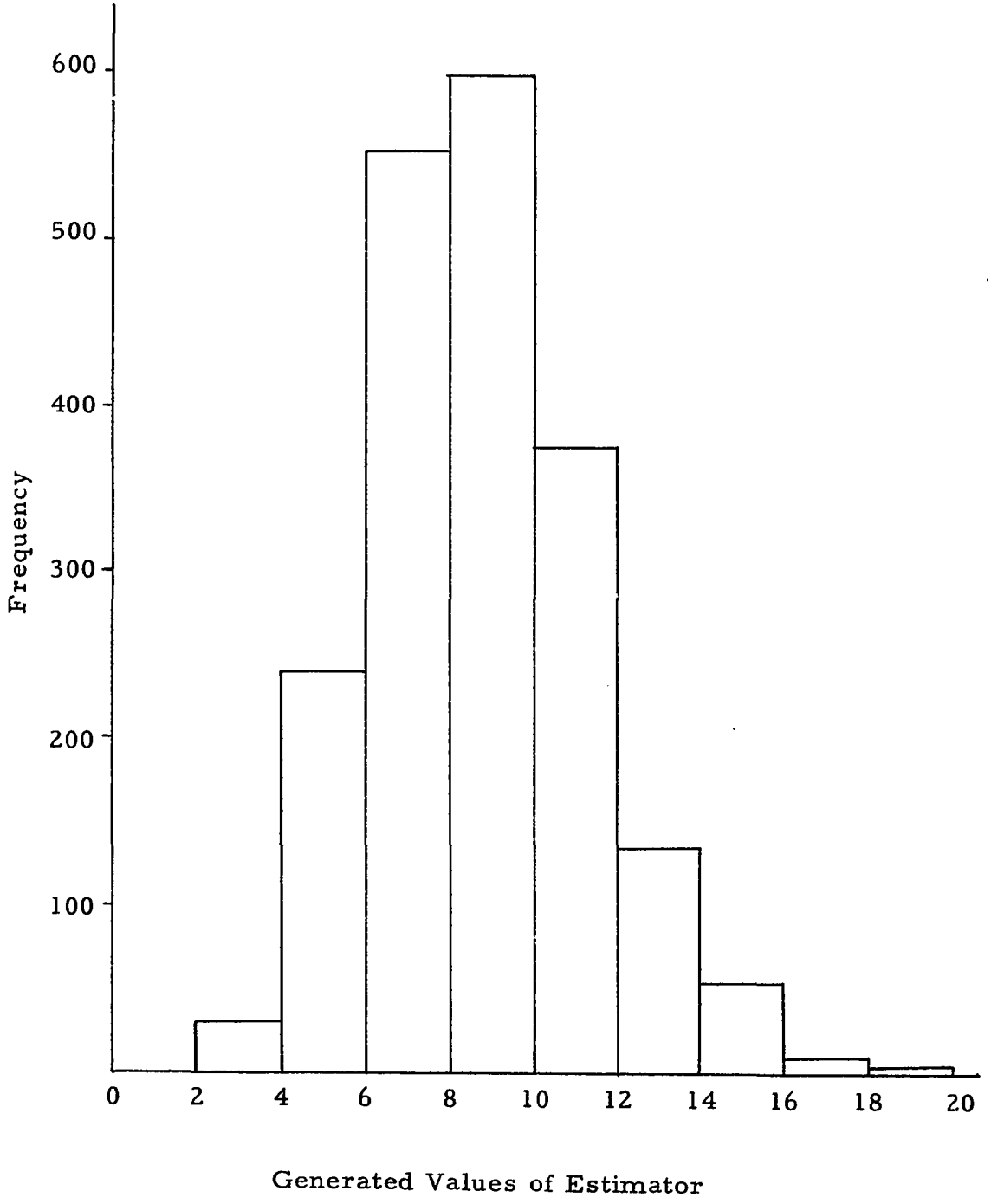

Figure 5.2(b). Histogram for $\bar{y}_{21}$ (Pop. 3)

Table 5.9(a). Frequency Distributions of Monte Carlo Generated Values of Estimators (2,000 Repetitions with n = 10 and n' = 30) for Population 1

Estimator Class	$\bar{y}_{21}$	$\bar{y}_{22}$	$\bar{y}_{23A}$	$\bar{y}_{23B}$	$\bar{y}_{24}$	$\bar{y}_{25}$	$\bar{y}_{2R}$
Less than 1	0	0	0	0	0	0	0
1- 1.99	4	4	3	4	4	4	1
2- 2.99	44	45	43	42	42	42	13
3- 3.99	87	91	75	80	76	76	55
4- 4.99	133	138	125	126	125	125	100
5- 5.99	160	155	151	155	149	148	154
6- 6.99	221	208	206	213	214	214	259
7- 7.99	238	228	220	228	220	221	279
8- 8.99	256	215	243	238	239	239	303
9- 9.99	209	211	218	225	231	231	279
10-10.99	199	169	191	187	189	185	200
11-11.99	141	162	173	165	171	175	140
12-12.99	99	103	101	100	97	95	99
13-13.99	87	92	89	89	90	90	51
14-14.99	57	62	65	62	66	67	34
15-15.99	31	53	46	40	40	41	17
16-16.99	17	28	27	25	25	24	8
17-17.99	6	16	10	8	8	9	6
18-18.99	7	10	7	8	8	8	1

Table 5.9(a) (Continued).

Estimator Class	$\bar{y}_{21}$	$\bar{y}_{22}$	$\bar{y}_{23A}$	$\bar{y}_{23B}$	$\bar{y}_{24}$	$\bar{y}_{25}$	$\bar{y}_{2R}$
19-19.99	2	4	3	1	2	2	1
20-20.99	1	2	3	3	3	3	0
21-21.99	1	2	1	1	1	1	0
22+	0	2	0	0	0	0	0

Table 5.9(b). Estimated Measures of Skewness ($\hat{\gamma}_1$) and Kurtosis* ($\hat{\gamma}_2$) for Distributions of Estimators Under Monte Carlo Simulation

Population/ Estimator	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\gamma}_1 = \hat{\mu}_3/(\hat{\mu}_2)^{\frac{3}{2}}$
Population 1				
$\bar{y}_{21}$	10.358	12.719	321.9	0.3816
$\bar{y}_{22}$	12.378	20.777	459.7	0.4773
$\bar{y}_{23A}$	11.170	13.976	374.3	0.3745
$\bar{y}_{23B}$	10.895	13.537	356.1	0.3764
$\bar{y}_{24}$	10.934	13.295	358.7	0.3677

*$\hat{\gamma}_2 = (\hat{\mu}_4/(\hat{\mu}_2)^2) - 3 \doteq 0$ in all the cases.

Table 5.9(b) (Continued).

Population/ Estimator	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\gamma}_1 = \hat{\mu}_3/(\hat{\mu}_2)^{\frac{3}{2}}$
$\bar{y}_{25}$	10.953	13.331	359.9	0.3679
$\bar{y}_{2R}$	7.141	7.4310	153.0	0.3893
Population 2				
$\bar{y}_{21}$	11.143	23.572	372.5	0.6337
$\bar{y}_{22}$	12.502	30.557	468.9	0.6912
$\bar{y}_{23A}$	11.950	25.834	428.4	0.6255
$\bar{y}_{23B}$	11.682	25.107	409.4	0.6289
$\bar{y}_{24}$	11.598	24.545	403.5	0.6214
$\bar{y}_{25}$	11.612	24.596	404.5	0.6214
$\bar{y}_{2R}$	10.468	22.483	328.7	0.6640
Population 3				
$\bar{y}_{21}$	6.321	7.4411	119.9	0.4664
$\bar{y}_{22}$	7.844	12.115	184.6	0.5517
$\bar{y}_{23A}$	6.760	8.024	137.1	0.4565
$\bar{y}_{23B}$	6.613	7.8336	131.2	0.4605
$\bar{y}_{24}$	6.500	7.5110	126.8	0.4533
$\bar{y}_{25}$	6.524	7.5264	127.7	0.4518
$\bar{y}_{2R}$	5.937	7.3742	105.7	0.5096
Population 4				
$\bar{y}_{21}$	5.145	4.9057	79.41	0.4204

Table 5.9(b) (Continued).

Population Estimator	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\mu}_4$	$\hat{\gamma}_1 = \hat{\mu}_3/(\hat{\mu}_2)^{\frac{3}{2}}$
$\bar{y}_{22}$	6.473	7.7713	125.7	0.4718
$\bar{y}_{23A}$	5.005	5.0071	75.16	0.4471
$\bar{y}_{23B}$	5.044	4.9437	76.32	0.4367
$\bar{y}_{24}$	5.174	4.9351	80.32	0.4193
$\bar{y}_{25}$	5.176	4.9311	80.37	0.4186
$\bar{y}_{2R}$	5.280	4.8510	83.63	0.3999

(2) Biases

From Table 5.10, we infer that the differences in bias as obtained by the two methods, Monte Carlo (β_M) and Asymptotic Expansion (β_A), are very small. For example in Population 1, the maximum value of $|\beta_M - \beta_A|/\bar{Y}$ is 0.00760 or 0.760 per cent. The remarkable closeness of the two sets of results is also shown in the other populations which have much smaller percentage differences. We note also the closeness of the expression $\beta_M - \beta_A$ for all estimators (except $\bar{y}_{2R}$) for Populations 1, 2 and 4. However, the ranking of the estimators with respect to absolute value of bias (from lowest to highest) may differ slightly between the two methods. Also, although one would normally assume that the results of the Monte Carlo simulation with 2,000 repetitions are

Table 5.10. Comparison between Monte Carlo (M) and Asymptotic Expansion (A) Estimates of Bias Under Scheme II for $(n, n') = (10, 30)$

Population/ Estimator	β_M	$\frac{\beta_M}{\bar{Y}} \times 100$	β_A	$\beta_M - \beta_A$	$\frac{\lvert\beta_M - \beta_A\rvert}{\bar{Y}} \times 100$
Population 1					
$\bar{y}_{21}$	-0.3168	-3.534	-0.2608	-0.0560	0.625
$\bar{y}_{22}$	-0.0681	-0.760	0.0	-0.0681	0.760
$\bar{y}_{23A}$	-0.0430	-0.480	0.0161	-0.0591	0.659
$\bar{y}_{23B}$	-0.1340	-1.495	-0.0762	-0.0578	0.645
$\bar{y}_{24}$	-0.0898	-1.002	-0.0298	-0.0600	0.670
$\bar{y}_{25}$	-0.0850	-0.948	-0.0253	-0.0597	0.666
$\bar{y}_{2R}$	-0.3845	-4.290	-0.3710	-0.0135	0.151
Population 2					
$\bar{y}_{21}$	-0.1625	-1.813	-0.1922	0.0297	0.331
$\bar{y}_{22}$	0.0301	0.336	0.0	0.0301	0.336
$\bar{y}_{23A}$	0.0867	0.967	0.0558	0.0309	0.345
$\bar{y}_{23B}$	0.0040	0.045	-0.0268	0.0308	0.345
$\bar{y}_{24}$	0.0154	0.017	-0.0129	0.0283	0.316
$\bar{y}_{25}$	0.0197	0.022	-0.0093	0.0290	0.324
$\bar{y}_{2R}$	-0.3831	-4.274	-0.3961	0.0130	0.145
Population 3					
$\bar{y}_{21}$	-0.2466	-2.751	-0.2437	-0.0029	0.032
$\bar{y}_{22}$	-0.0085	-0.095	0.0	-0.0085	0.095

Table 5.10 (Continued).

Population/ Estimator	β_M	$\frac{\beta_M}{\bar{Y}} \times 100$	β_A	$\beta_M - \beta_A$	$\frac{\lvert\beta_M - \beta_A\rvert}{\bar{Y}} \times 100$
$\bar{y}_{23A}$	-0.0042	-0.047	0.0021	-0.0063	0.070
$\bar{y}_{23B}$	-0.0852	-0.951	-0.0798	-0.0054	0.060
$\bar{y}_{24}$	-0.0776	-0.868	-0.0657	-0.0119	0.133
$\bar{y}_{25}$	-0.0630	-0.703	-0.0481	-0.0149	0.166
$\bar{y}_{2R}$	-0.1531	-1.708	-0.1069	-0.0462	0.515
Population 4					
$\bar{y}_{21}$	-0.0241	-0.269	-0.0072	-0.0169	0.189
$\bar{y}_{22}$	-0.0103	-0.115	0.0	-0.0103	0.115
$\bar{y}_{23A}$	-0.2775	-3.060	-0.2651	-0.0124	0.138
$\bar{y}_{23B}$	-0.1928	-2.151	-0.1792	-0.0136	0.152
$\bar{y}_{24}$	-0.0290	-0.324	-0.0136	-0.0154	0.172
$\bar{y}_{25}$	-0.0304	-0.339	-0.0122	-0.0182	0.203
$\bar{y}_{2R}$	-0.0805	-0.898	-0.0405	-0.0400	0.446

more reliable, one should note that in the example above the true value of $\beta(\bar{y}_{22})$ is 0 .

We note from Table 5.11 that the value of β^2/σ^2 is less than 0.01 for all estimators and populations except $\bar{y}_{2R}$ (Populations 1 and 2) and $\bar{y}_{23A}$ (Population 4). Further, for these exceptions, β^2/σ^2 is less than 0.04. Therefore, one might ignore the effect of the bias on the MSE of the estimators for (n, n') = (10, 30) in the populations studied.

Table 5.11. Comparison between Monte Carlo (M) and Asymptotic Expansion (A) Estimates of MSE Under Scheme II for $(n, n') = (10, 30)$

Population/ Estimator	σ_M^2	$(\beta_M)^2$	$\frac{(\beta_M)^2}{\sigma_M^2}$	MSE(M)	MSE(A)	$\frac{MSE(A)}{MSE(M)}$
Population 1						
$\bar{y}_{21}$	10.358	0.100	0.0097	10.458	10.715	1.025
$\bar{y}_{22}$	12.378	0.005	0.0004	12.383	12.682	1.024
$\bar{y}_{23A}$	11.170	0.002	0.0002	11.172	11.462	1.026
$\bar{y}_{23B}$	10.895	0.018	0.0017	10.913	11.195	1.026
$\bar{y}_{24}$	10.934	0.008	0.0007	10.942	11.225	1.026
$\bar{y}_{25}$	10.953	0.007	0.0006	10.960	11.225	1.024
$\bar{y}_{2R}$	7.141	0.148	0.0207	7.289	7.298	1.001
Population 2						
$\bar{y}_{21}$	11.143	0.026	0.0023	11.169	12.011	1.075
$\bar{y}_{22}$	12.502	0.001	0.0001	12.503	13.316	1.065
$\bar{y}_{23A}$	11.950	0.008	0.0007	11.958	12.790	1.070
$\bar{y}_{23B}$	11.682	0.000	0.0000	11.682	12.516	1.071
$\bar{y}_{24}$	11.598	0.000	0.0000	11.598	12.401	1.069
$\bar{y}_{25}$	11.612	0.000	0.0000	11.612	12.401	1.068
$\bar{y}_{2R}$	10.468	0.147	0.0140	10.615	10.947	1.031
Population 3						
$\bar{y}_{21}$	6.321	0.061	0.0097	6.382	6.884	1.079
$\bar{y}_{22}$	7.844	0.000	0.0000	7.844	8.468	1.080

Table 5.11 (Continued).

Population/ Estimator	σ_M^2	$(\beta_M)^2$	$\frac{(\beta_M)^2}{\sigma_M^2}$	MSE(M)	MSE(A)	$\frac{MSE(A)}{MSE(M)}$
$\bar{y}_{23A}$	6.760	0.000	0.0000	6.760	7.269	1.075
$\bar{y}_{23B}$	6.613	0.007	0.0011	6.620	7.127	1.077
$\bar{y}_{24}$	6.500	0.006	0.0009	6.506	6.977	1.072
$\bar{y}_{25}$	6.524	0.004	0.0006	6.528	6.977	1.069
$\bar{y}_{2R}$	5.937	0.023	0.0039	5.960	6.200	1.040
Population 4						
$\bar{y}_{21}$	5.145	0.001	0.0002	5.146	5.722	1.119
$\bar{y}_{22}$	6.473	0.000	0.0000	6.473	7.008	1.082
$\bar{y}_{23A}$	5.005	0.077	0.0154	5.082	5.610	1.104
$\bar{y}_{23B}$	5.044	0.037	0.0073	5.081	5.635	1.109
$\bar{y}_{24}$	5.174	0.001	0.0002	5.175	5.763	1.114
$\bar{y}_{25}$	5.176	0.001	0.0002	5.177	5.763	1.113
$\bar{y}_{2R}$	5.280	0.006	0.0011	5.286	5.756	1.089

(3) MSE

In Table 5.11 the values of the MSE's of the various estimators obtained by the two methods, Monte Carlo [MSE(M)] and Asymptotic Expansion [MSE(A)], are given. For consistency, the ratio MSE(A)/MSE(M) should be around unity. We find that for all four populations, this ratio lies between 1.001 for $\bar{y}_{2R}$ (Population 1) and 1.119 for $\bar{y}_{21}$

(Population 4); i.e. taking MSE(M) as a standard, the discrepancies range from 0.1 per cent to 11.9 per cent. This is, on the whole, quite satisfactory for the moderately small values of n and n' considered, and also for the relatively small values of MSE's involved (especially in Population 4).

As in the case of the bias, the order of ranking of the MSE's is essentially the same in both methods. There is only one permutation in ranking for all four populations. However, one should note that the differences in magnitude of the MSE's of the estimators are in general rather small.

VI. OPTIMAL STRATIFICATION

In Chapter I, we outlined briefly the application of two-phase sampling to optimal stratification. We elaborate further in this chapter.

Using the cum$\sqrt{f}$ method in the single phase case, Serfling [50] has derived (ignoring the f.p.c.) an approximation to the minimum value of $V(\bar{y}_{st})$. For ease of presentation, we summarize his derivation in the case where the stratification variable is the same as the variable of interest, Y. We recall from Dalenius and Hodges [11, 12] that if we represent the p.d.f. of Y by f(y), $a \leq y \leq b$ (f(y) = 0 elsewhere), then the optimal stratification boundaries (O.S.B.) z_r, r = 1, 2, ..., L-1 are determined by

$$A_r = \int_{z_{r-1}}^{z_r} \sqrt{f(y)}\, dy = K/L \qquad (6.1)$$

where L = number of strata and,

$$K = \int_a^b \sqrt{f(y)}\, dy .$$

For sufficiently large L, f(y) can be approximated by its "mean value" C_r within the rth stratum. Approximations for the weight, variance and "cum $\sqrt{f}$" of this rth stratum are given by;

$$W_r \doteq C_r(z_r - z_{r-1}) \tag{6.2a}$$

$$S_r^2 \doteq (z_r - z_{r-1})^2/12 \tag{6.2b}$$

$$A_r \doteq \sqrt{C_r}\,(z_r - z_{r-1}) \tag{6.2c}$$

where A_r is as defined in (6.1).

We recall that, using Neyman's allocation and ignoring the f.p.c.,

$$V(\bar{y}_{st}) = \frac{1}{n}(\Sigma W_r S_r)^2 \tag{6.3}$$

Hence, from (6.2a), ..., (6.3), minimizing $V(\bar{y}_{st})$ is equivalent to minimizing $\sum_{r=1}^{L} A_r^2$. Since $\sum_{r=1}^{L} A_r = K$ irrespective of the choice of the stratification boundaries, the optimal choice for the boundaries is given by $A_r = K/L$. Then it is easily seen that

$$V_{min}(\bar{y}_{st}) \doteq K^4/12nL^2 \tag{6.4}$$

We next consider the accuracy of Serfling's approximation for the minimum variance of $\bar{y}_{st}$ in single phase sampling. We compare in Table 6.1 the exact value of $n\, V_{min}(\bar{y}_{st})$ with $K^4/12L^2$ for each of the following distributions:

(1) the right-triangular

$$f_1(y) = 2(1-y), \qquad 0 < y < 1$$

$$= 0, \qquad \text{otherwise}$$

(2) the negative exponential

$$f_2(y) = e^{-y}, \qquad y > 0$$

$$= 0, \qquad \text{otherwise}$$

(3) the gamma distribution with parameter 2

$$f_3(y) = ye^{-y}, \qquad y > 0$$

$$= 0, \qquad \text{otherwise}$$

(4) the half-normal

$$f_4(y) = \left(\frac{2}{\pi}\right)^{\frac{1}{2}} e^{-y^2/2}, \qquad y \geq 0$$

$$= 0, \qquad \text{otherwise}$$

The exact minimum values based on evaluation of $(\Sigma W_r S_r)^2$ given for f_1, f_2 and f_3 are taken from [14] while those for f_4 are from [39]. The f_4 minima are based on equal allocation; that is, $\frac{1}{n}\Sigma W_r^2 S_r^2$ is evaluated for the O.S.B.

From Table 6.1, it seems that Serfling's approximation is fairly satisfactory for $L > 3$. Of the four populations considered, the approximation is best for the right-triangular distribution while it is poorest for the gamma distribution. However, as L increases, the relative error for the gamma distribution declines very rapidly.

So far Y has been assumed to be a continuous random variable. But it seems reasonable to assume that the "cum $\sqrt{f}$" method can be adapted to the discrete case [9]. Thus, assume that the variable Y takes the values $y_1, y_2, \ldots, y_T$ with relative frequencies $P_1, P_2, \ldots, P_T$ where $P_i = N_i/N$ $(i = 1, 2, \ldots, T)$ and $\sum_{i=1}^{T} P_i = 1$. We also assume without loss of generality that $y_1 < y_2 < \ldots < y_T$. (We note that this representation of Y is equivalent to the Y values being grouped into T classes, of size N_i, for $i = 1, 2, \ldots, T$).

The (single phase) approach to the discrete case would thus follow the same lines as for the continuous case with sums replacing integrals. We assume for simplicity that all the class intervals have equal length;

Table 6.1. Comparison of Actual Minimum of $(\sum_{r=1}^{L} W_r S_r)^2$ with $K^4/12L^2$ for Certain Functions

Density function	Quant.	L = 2	3	4	5
f_1	Exact				
	Min* (1)	0.01505	0.00688	0.00393	0.00254
	$K^4/12L^2$ (2)	0.01645	0.00732	0.00412	0.00263
	% Error (3)	9.2	6.4	4.8	3.5
f_2	(1)	0.2855	0.1332	0.0768	0.0500
	(2)	0.3333	0.1481	0.0833	0.0533
	(3)	16.8	11.2	8.6	6.6
f_3	(1)	0.6370	0.3075	0.1804	0.1185
	(2)	0.8222	0.3656	0.2056	0.1316
	(3)	29.1	19.0	14.0	11.0
f_4	(1)	0.1096	0.0516	0.0304	-
	(2)	0.1309	0.0582	0.0327	-
	(3)	19.3	12.8	7.5	-

without loss of generality, we take this to be unity. We also assume that the partition into L strata has all members of one class in a single stratum. Similar to the continuous case, we define

$$K = \sum_{i=1}^{T} \sqrt{P_i} \tag{6.5}$$

The "cum $\sqrt{f}$" rule for the discrete case is then

$$\sum_{j=1}^{z_1} \sqrt{P_j} = \sum_{j=z_1-1}^{z_2} \sqrt{P_j} = \ldots = K/L \tag{6.6}$$

where the z_i's are the O. S. B. For the discrete case, analogous expressions to (6.2a), ..., (6.2c) are used.

If the ith class interval has length u_i, then (6.5) becomes:

$$K = \sum_{i=1}^{T} \Sigma \sqrt{(u_i P_i')} \tag{6.7}$$

If it is assumed that N is known but the N_i are unknown, a double sampling scheme may be applied. Thus, we select a preliminary sample of size n' resulting in the class frequencies n_i' ($i = 1, 2, \ldots, T$) with $\sum_{i=1}^{T} n_i' = n'$, and $P_i' = n_i'/n'$.

We then use the adaptation of the "cum $\sqrt{f}$" method to the discrete random variable (outlined above) to determine the optimal boundaries of the first phase sample. We define

$$K = \sum_{i=1}^{T} \sqrt{P'_i} \tag{6.8}$$

and denote by $P_{s_{n'}}$, a specific preliminary sample of size n'. Hence given $P_{s_{n'}}$, the O.S.B. are determined by choosing the boundary points z_i $(i = 1, 2, \ldots, L-1)$ such that $\sum_{j=z_i+1}^{z_{i+1}} \sqrt{P'_j}$ is constant.

We recall that in two-phase sampling

$$V_1(\bar{y}_{std}) = E[V_1(\bar{y}_{std}|P_{s_{n'}})] + V_1[E(\bar{y}_{std}|P_{s_{n'}})] \tag{6.9}$$

where E refers to all possible samples of size n'. But

$$V_1(\bar{y}_{std}|P_{s_{n'}}) = \frac{1}{n}\left(\sum_{r=1}^{L} W'_r S'_r\right)^2 - \frac{1}{n'}\sum_{r=1}^{L} W'_r (S'_r)^2 \tag{6.10}$$

where W'_r and $(S'_r)^2$ are the preliminary sample weight and the preliminary sample mean square for the rth stratum. If we assume that n' is sufficiently large so that minimizing the first term on the r.h.s. of (6.10) is adequate to determine the O.S.B. for the preliminary sample, then Serfling's approximation leads to

$$V_1(\bar{y}_{std}|P_{s_{n'}}) = \frac{\bar{K}^2}{12nL^2} - \frac{\bar{K}^2}{12nL^2}(b-a) \qquad (6.11)$$

The first term on the r.h.s. of (6.11) comes from (6.4) while the second term is obtained by making the appropriate substitutions from (6.2a) and (6.2b) in the second term on the r.h.s. of (6.10). We note that for (6.10) we have assumed that $P_i' \neq 0$. Even if $P_i' = 0$, we can always "stretch" the first and last strata to accommodate the extreme points of the population.

Ignoring the first phase f.p.c., (6.9) becomes:

$$V_1(\bar{y}_{std}) = E[\frac{\bar{K}^4}{12nL^2} - \frac{\bar{K}^2}{12n'L^2}(b-a)] + \frac{S^2}{n'} \qquad (6.12)$$

Expanding $\bar{K}$ in a Taylor series about $(P_1' = P_1, \ldots, P_T' = P_T)$, we obtain

$$E(\bar{K}^4) \doteq K^4 + \frac{K^2}{n'}[\frac{3T}{2} - K^2 - \frac{1}{2}\{(\sum_{i=1}^{T}\sqrt{P_i})(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}})\}] \qquad (6.13a)$$

$$E(\bar{K}^2) \doteq K^2 + \frac{1}{4n'}[T - \{(\sum_{i=1}^{T}\sqrt{P_i})(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}})\}] \qquad (6.13b)$$

Also, since we are using 1 unit class intervals, $b-a = T-1$. Hence, (6.12) can be written as:

$$V_1(\bar{y}_{std}) = \frac{K^2}{12L^2}\left(\frac{K^2}{n} - \frac{T-1}{n'}\right) + \frac{S^2}{n'}$$

$$+ \frac{K^2}{12nn'L^2}\left[\frac{3T}{2} - K^2 - \frac{1}{2}\left\{\left(\sum_{i=1}^{T}\sqrt{P_i}\right)\left(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}}\right)\right\}\right]$$

$$- \frac{(T-1)}{12L^2}\left[\frac{1}{4(n')^2}\left\{T - \left(\sum_{i=1}^{T}\sqrt{P_i}\right)\left(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}}\right)\right\}\right] \qquad (6.14)$$

We note that if we consider the rather unrealistic situation of the O.S.B. (for the population of N elements) being known prior to any sampling and the preliminary sample being used only to obtain estimates of the unknown stratum weights, then from the equation below (12.6) in Cochran [9] with $\lambda_h = S_h$ and $g' = 1$

$$V_2(\bar{y}_{std}) = \left(\sum_{i=1}^{L} W_i S_i\right)^2/n + \sum_{i=1}^{L} W_i(\bar{Y}_i - \bar{Y})^2/n'$$

$$+ \frac{1}{nn'}\left[\sum_{i=1}^{L} W_i S_i^2 - \left(\sum_{i=1}^{L} W_i S_i\right)^2\right] \qquad (6.15)$$

We note that (6.15) holds for any choice of stratum boundaries. We recall that

$$\frac{S^2}{n'} \doteq \frac{1}{n'} \sum_{i=1}^{L} \{W_i S_i^2 + W_i(\bar{Y}_i - \bar{Y})^2\} \tag{6.16}$$

Thus from (6.15) and (6.16),

$$V_2(\bar{y}_{std}) = \frac{(\sum_{i=1}^{L} W_i S_i)^2}{n} - \frac{1}{n'} \sum_{i=1}^{L} W_i S_i^2 + \frac{S^2}{n'}$$

$$+ \frac{1}{nn'} [\sum_{i=1}^{L} W_i S_i^2 - (\sum_{i=1}^{L} W_i S_i)^2] \tag{6.17}$$

Using Serfling's approximations and noting that $b-a = T-1$, we can write

$$V_2(\bar{y}_{std}) = \frac{K^2}{12L^2}(\frac{K^2}{n} - \frac{T-1}{n'}) + \frac{S^2}{n'}$$

$$+ \frac{1}{nn'} [\frac{K^2}{12L^2}(T-1-K^2)] \tag{6.18}$$

From (6.14) and (6.18),

$$V_1(\bar{y}_{std}) - V_2(\bar{y}_{std}) = \frac{K^2}{12nn'L^2}\left[\frac{T}{2} + 1 - \frac{1}{2}\left(\sum_{i=1}^{T}\sqrt{P_i}\right)\left(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}}\right)\right]$$

$$- \frac{T-1}{12(n')^2L^2}\left[\frac{1}{4}\left\{T - \left(\sum_{i=1}^{T}\sqrt{P_i}\right)\left(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}}\right)\right\}\right] \qquad (6.19)$$

But, from the Cauchy-Schwarz inequality,

$$\left(\sum_{i=1}^{T}\sqrt{P_i}\right)\left(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}}\right) \geq T^2 \qquad (6.20)$$

Hence to $O(\frac{1}{nn'})$,

$$V_1(\bar{y}_{std}) - V_2(\bar{y}_{std}) \leq 0 \qquad (6.21)$$

Hence to this order of approximation, it would seem that our approach leads to a variance which is not larger than that given by Cochran. If we consider terms to $O(\frac{1}{(n')^2})$, we rewrite (6.19) as:

$$V_1(\bar{y}_{std}) - V_2(\bar{y}_{std}) = \frac{1}{12L^2n'}\left[\left(\frac{T^2 - T + a}{2}\right)\left\{\frac{T-1}{2n'}\right.\right.$$

$$\left.\left. - \frac{K^2}{n}\left(1 - \frac{2}{T^2 - T + a}\right)\right\}\right] \qquad (6.22)$$

where from (6.20), we put

$$\left(\sum_{i=1}^{T}\sqrt{P_i}\right)\left(\sum_{i=1}^{T}\frac{1}{\sqrt{P_i}}\right) = T^2 + a \quad ,$$

a being a non-negative quantity. Hence since $T^2 - T + a \geq 0$ for $T \geq 1$,

$$V_1(\bar{y}_{std}) - V_2(\bar{y}_{std}) \leq 0$$

if

$$\frac{n}{n'} < \frac{2K^2}{T-1}\left(1 - \frac{2}{T^2 - T + a}\right) \tag{6.23}$$

We note, however, that to $O(\frac{1}{n'})$ the two methods lead to the same variance, conditional upon our assumptions being valid.

We note that for L small, a practical method of finding the O.S.B. is to use a trial and error approach (for either the total population or the preliminary sample). This would involve partitioning the population into L strata in all possible ways and deducing by actual calculations that partition which gives the minimum variance. The work involved is substantially reduced if it is assumed that stratum 1 consists of those units with $Y = y_1, y_2, \ldots, y_i$ (say), stratum 2 of $Y = y_{i+1}, \ldots,$ etc. Alternatively, one might use those boundaries obtained by application of the "cum $\sqrt{f}$" method as a starting point in determining the actual O.S.B.

In the preceding discussion, we have considered the case where the selection of the second phase sample is based on Neyman's allocation (or what is roughly equivalent, equal allocation). Clearly, a parallel argument can be set forth for the case where the second phase sample is selected by proportional allocation.

VII. ESTIMATION OF DOMAIN MEANS

A. Specification of Problem

In certain surveys, the primary objective is to estimate stratum (or domain) means $\bar{Y}_1, \ldots, \bar{Y}_L$. In planning such surveys, one might specify the desired precision of each of the estimated stratum means, $\bar{y}_k$, $k = 1, 2, \ldots, L$. Then, how should a fixed total sample size be allocated among the various strata? In a recent paper, Chaddha et al. [4] discuss this allocation problem when the estimated stratum means are to have their variances in a given ratio i.e. $\frac{V(\bar{y}_k)}{V_k} = \frac{\sigma_k^2}{V_k n_k} \doteq$ const. for all k, where V_k is specified by the investigator. More precisely, the allocation problem is

$$\underset{\underset{\sim}{n} \in G}{\text{Min}} \left[\underset{k}{\text{Max}} \{\varphi_k\} \right] \tag{7.1}$$

where $\underset{\sim}{n} = (n_1, n_2, \ldots, n_L)$

$$G = \{\underset{\sim}{n}: \; n_k \text{ an integer}, \; 1 \leq n_k \leq N_k ,$$

$$k = 1, 2, \ldots, L \text{ and } \sum_{k=1}^{L} n_k = n\}$$

$$\varphi_k = \frac{\sigma_k^2}{V_k n_k}$$

and N_k = total number of units in the kth stratum.

Chaddha et al. have proposed three methods for solving this problem. Two of these yield exact integer-valued optimal solutions. We note that Chaddha et al. assume that the units belonging to each stratum are identifiable before any sampling is done.

We are interested in estimating the means of the D domains comprising the population. It is assumed that there is no prior identification of the units with respect to their domains. Therefore, a preliminary random sample of size n' is selected and the units are classified by domains. Then, within each domain, a simple random subsample of the first phase units is drawn and the domain means are estimated. In the preliminary sample, n'_j units are assumed to be members of the jth domain ($\sum_{j=1}^{D} n'_j = n'$); a subsample of size n_j is chosen from these n'_j units using an unequivocal sampling rule (S.R.). Thus, the n_j are random variables because they are functions of the n'_j.

Given the values for the n'_j ($j = 1, 2, \ldots, D$) and n, one might choose the n_j to minimize the maximum (over the domains) value of $\frac{\sigma_j^2}{V_j n_j}$ where $\frac{\sigma_j^2}{n_j}$ is the conditional variance of the estimated domain mean. We note that $1 \leq n_j \leq n'_j$, since we assume* that $n'_j \neq 0$, $j = 1$,

---

*If $n'_j = 0$ for any domain, we assume, for convenience, that the preliminary sample is re-drawn. Of course, for most practical situations n' will be chosen sufficiently large that $P(n'_j = 0)$ is negligible.

2, ..., D. Also the optimal values of the n_j should be integers. To choose n' and n, one might find those values which satisfy the budget constraint

$$C^* = c'n' + cn \tag{7.2}$$

and minimize (with respect to n', n) the maximum (over the domains) value of $E(\frac{\sigma_j^2}{V_j n_j})$. Here

c' = the cost per unit of selecting and identifying the first phase units,

and c = the cost per unit of selecting and measuring the second phase units.

B. The Complete Sampling Rule (S. R.)

We first give the optimal (but not necessarily integer-valued) allocation of a fixed sample size, n, among the domains given a specified preliminary sample. To ensure an integer-valued solution, some modifications of the procedure are required. These are taken up after the basic results are presented. Also we defer until later proofs of our results.

We first consider the simple case $D = 2$ and then generalize: Let $a_j = \sigma_j^2/V_j$ and define

$$n_j^* = \frac{a_j n}{\sum_{j=1}^{2} a_j} \quad \text{for} \quad j = 1, 2 \; . \tag{7.3}$$

If $n_j^* \leq n_j'$, $j = 1, 2$, the allocation is

$$n_1 = n_1^* \; , \qquad n_2 = n_2^* \; .$$

However, if one of the values, n_1^* (say) is greater than n_1' , then the maximum value of n_1 will be n_1' . This reasoning leads to the following complete S.R.:

$$\text{take} \begin{cases} n_1 = n_1^* \text{ if } (n + n_1' - n') \leq n_1^* \leq n_1' \\ n_1 = n_1' \text{ if } n_1^* > n_1' \\ n_1 = n + n_1' - n' \text{ if } n_1^* < (n + n_1' - n') \end{cases}$$

$$n_2 = n - n_1$$

If $D > 2$, the S.R. becomes more complicated. For general D , we proceed as in the following algorithm:

(1) Define $$n_i^* = \frac{a_i n}{\sum_{j=1}^{D} a_j}, \quad (i = 1, 2, \ldots, D) . \tag{7.4}$$

Proceed to (2).

(2) If $n_i' \geq n_i^*$, for all i, stop. The allocation is

$$n_i = n_i^* \qquad (i = 1, 2, \ldots, D) .$$

If $n_i' < n_i^*$ for at least one i, proceed to (3).

(3) Assume, without loss of generality, that for some k

$$n_i' \geq n_i^* \qquad (i = 1, 2, \ldots, k)$$

but

$$n_i' < n_i^* \qquad (i = k + 1, \ldots, D)$$

Take

$$n_i = n_i' , \qquad (i = k + 1, k + 2, \ldots, D) \tag{7.5}$$

Proceed to (4).

(4) Put $\bar{n} = n - \sum_{k+1}^{D} n_i$ and define

$$n_i^{**} = \frac{a_i \bar{n}}{\sum_{j=1}^{k} a_j} , \quad (i = 1, 2, \ldots, k) .$$

Proceed to (5).

(5) If $n_i' \geq n_i^{**}$, for $i = 1, 2, \ldots, k$, stop. The remaining allocation is

$$n_i = n_i^{**} \qquad (i = 1, 2, \ldots, k)$$

If $n_i' < n_i^{**}$ for at least one i in the set $i = 1, 2, \ldots, k$, go to (6).

(6) Again assume, without loss of generality, that

$$n_i' \geq n_i^{**} \qquad (i = 1, 2, \ldots, k^*)$$

$$n_i' < n_i^{**} \qquad (i = k^* + 1, \ldots, k)$$

Then, take

$$n_i = n_i' \qquad (i = k^* + 1, \ldots, k)$$

and define $\bar{\bar{n}} = \bar{n} - \sum_{i=k^*+1}^{k} n_i$ and proceed as in (4) with $\bar{\bar{n}}$ replacing $\bar{n}$, and n_i^{***} replacing n_i^{**}.

(7) Continue the process until the following is obtained:

$$n_i' \geq n_i^* \ldots^* \quad , \quad \text{for all remaining } i$$

Then the remaining allocation is

$$n_i = n_i^* \ldots * \quad , \quad \text{for all remaining } i \ .$$

It is easy to show that the process terminates after a finite number of steps.

It is easy to show that the allocation is the "minimax solution" . We note that if

$$\frac{a_j}{n_j} = \text{const.}, \quad j = 1, 2, \ldots, D \tag{7.6}$$

the minimax solution is obtained. If $n_j' \geq n_j^*$, $j = 1, 2, \ldots, D$, then (7.6) is automatically satisfied by the allocation rule.

However, if $n_j' < n_j^*$ for $j = k + 1, \ldots, D$, we note that

$$\frac{a_j}{n_j'} = \frac{a_i}{n_i^*}$$

for any j ($j \in \{k + 1, \ldots, D\}$) and for any i ($i \in \{1, 2, \ldots, k\}$) . Hence if

$$n_j^* \leq n_j^{**} \leq \cdots \leq n_j^{*\cdots *} , \tag{7.7}$$

we can conclude that $\max_j(\frac{a_j}{n_j})$ comes from the set $\{k+1, \ldots, D\}$. Since $\max_j(\frac{a_j}{n_j})$ for $j = k+1, \ldots, D$ is as small as it can be, we will have the "minimax" solution.

To show that $n_j^* \leq n_j^{**}$, consider

$$n_j^{**} - n_j^* = a_j\left[\frac{\bar{n}}{\sum_{j=1}^{k} a_j} - \frac{n}{\sum_{j=1}^{D} a_j}\right]$$

$$= \frac{a_j\left[n \sum_{k+1}^{D} a_j - \sum_{k+1}^{D} n_i \sum_{1}^{D} a_j\right]}{\left(\sum_{j=1}^{k} a_j\right)\left(\sum_{1}^{D} a_j\right)} \tag{7.8}$$

We need only show that

$$\left[n \sum_{k+1}^{n} a_j - \sum_{k+1}^{D} n_i \sum_{1}^{D} a_j\right] \geq 0$$

and then the result follows. We note that from (7.4) and (7.5)

$$\sum_{k+1}^{D} n_j < \frac{\sum_{k+1}^{D} a_j n}{\sum_{1}^{D} a_j} .$$

Hence the numerator of (7.7) is non-negative and

$$n_j^{**} \geq n_j^{*} .$$

Similarly, it can be shown that

$$n_j^{**} \leq n_j^{***} \leq \ldots \leq n_j^{*} \ldots * .$$

The allocation obtained by proceeding as described above will not, in general, provide integer-valued n_j. To obtain integer-valued solutions, the complete S.R. outlined above needs only slight modifications. We start with Step 1 in our algorithm but replace n_i^{*} with the integer-valued solutions n_i^{o} obtained from (7.1). We modify Steps 2 and 3 by replacing n_i^{*} with n_i^{o}. In Step 4, we again obtain integer-valued solutions, n_j^{oo}, to the minimax problem at this stage, and so on. We illustrate this by giving the complete S.R. for $D = 2, 3$:

(1) For $D = 2$, the complete S.R. is the same as that given below (7.3) with n_1^{o} replacing n_1^{*}.

(2) For $D = 3$, the complete S.R. is:

(a) if $1 \leq n_j^o \leq n_j'$, $j = 1, 2, 3$

take $n_1 = n_1^o$

$n_2 = n_2^o$

$n_3 = n - n_1 - n_2 \; (= n_3^o)$.

(b) If $n_1^o > n_1'$,

then take $n_1 = n_1'$

$$\begin{cases} n_2 = n_2^{oo} \text{ if } (n + n_2' - n') \leq n_2^{oo} \leq n_2' \\ n_2 = n_2' \text{ if } n_2^{oo} > n_2' \\ n_2 = n + n_2' - n' \text{ if } n_2^{oo} < (n + n_2' - n') \end{cases}$$

$n_3 = n - n_1 - n_2$.

(c) If $n_j^o > n_j'$ (for $j = 1, 2$, say),

take $n_1 = n_1'$

$$n_2 = n_2'$$

$$n_3 = n - n_i - n_2 .$$

E. Selection of Values for n' and n

In the preceding discussion, it has been assumed that both n' and n are fixed but unknown quantities. Assuming a given budget, and the linear cost function, (7.2), several choices of the pair (n', n) are possible. Thus, the objective is to find that value of (n', n) which minimizes the maximum (over the domains) value of $E(a_j/n_j)$. (The expectation operator refers to repeated selection of preliminary samples of size n' with the n_j determined from the n_j' by an appropriate S. R.). To accomplish this, we may consider a sequence of "trial" values of n' (the corresponding values of n are determined by the budget restriction). For each value of n', the maximum value of $E(a_j/n_j)$ is ascertained; and, finally, the optimal value of n' is chosen among the sequence of "trial" values of n'. Thus, for a given (trial) value of n', one must evaluate $E(n_j^{-1})$ for $j = 1, 2, \ldots, D$. However, this is difficult because the sampling rule is complex, and (for $D > 2$) tables of "multinomial" probabilities will be needed. Since tables of the relevant probabilities are not available, it is infeasible, without a computer, to determine the value of $E(n_j^{-1})$ if $D > 2$. Hence, we investigate approximations.

First it may be sufficient to evaluate $E(n_j^{-1})$ for only one (or a subset) of the D domains. This will be true if one (or a subset) of the

domains yields the maximum value of $E(a_j/n_j)$ irrespective of the value of n'. For example, $a_1 = a_2$ and $\pi_1 < 0.5$, domain 1 should always yield the maximum value (see Tables 7.2(a) and (b)).

Second, we recall from the discussion in Section B that the random variables n_j assumes the values $n_j^!$, n_j^0, ..., $n_j^{0\ldots0}$ where $n_j^0 \leq n_j^{00} \leq \ldots \leq n_j^{0\ldots0}$. Hence it can easily be shown that

$$E\left(\frac{1}{n_j}\right) \leq \frac{1}{n_j^!} \sum_{x=1}^{n_j^0 - 1} P(X=x) + \frac{1}{n_j^0} \sum_{x=n_j^0}^{n'-1} P(X=x) \qquad (7.9)$$

where, assuming that sampling is with replacement,

$$P(X=x \mid n', \pi_j) = \frac{\binom{n'}{x} \pi_j^x (1-\pi_j)^{n'-x}}{1 - P(X=0) - \sum_{j=0}^{D-2} P(X = n'-j)} ,$$

and $\pi_j = N_j/N$. In many practical situations, as previously implied, $P(X = 0)$ and $\sum_{j=0}^{D-2} P(X=n'-j)$ are sufficiently small that they may be neglected.

We also recall that

$$E\left(\frac{1}{n_j}\right) \geq \frac{1}{E(n_j)} \qquad (7.10)$$

Hence, putting

$$E^*(n_j) = \frac{1}{n_j'} \sum_{x=1}^{n_j^0 - 1} P(X=x) + \frac{1}{n_j^0} \sum_{x=n_j^0}^{n'-1} P(X=x) ,$$

we have, from (7.9) and (7.10),

$$\frac{1}{E(n_j)} \leq E(\frac{1}{n_j}) \leq E^*(n_j) \tag{7.11}$$

This provides an upper and a lower bound for $E(\frac{1}{n_j})$.

For the reasons cited above, it is difficult to evaluate $E(n_j)$ if a computer is not available. However, the upper bound for $E(n_j^{-1})$ can be evaluated with only the aid of binomial tables. Further, a crude approximation for $E(n_j)$ is suggested below.

We also note that we can use the usual Taylor series approximation for $E(\frac{1}{n_j})$, namely

$$E(\frac{1}{n_j}) \doteq \frac{1}{E(n_j)} [1 + C_{n_j}^2]$$

$$= \frac{1}{[E(n_j)]^3} [E(n_j^2)] \tag{7.12}$$

where C_{n_j} is the coefficient of variation of n_j. However, (7.12) is of little practical value (except possibly for $D = 2$) unless approximations for $E(n_j)$ and $E(n_j^2)$ are used. One possibility has been proposed by Sedransk [49; 3.22]; for example, for $D = 3$,

$$E(n_1) = \begin{cases} E(n_1') & \text{if } E(n_1') \leq n_1^* \\ n^* & \text{if } E(n_1') > n_1^* \end{cases}$$

$$E(n_2) = \begin{cases} E(n_2') & \text{if } E(n_2') \leq (n - En_1)\left(\frac{a_2}{a_2 + a_3}\right) \\ (n - En_1)\left(\frac{a_2}{a_2 + a_3}\right) & \text{if } E(n_2') > (n - En_1)\left(\frac{a_2}{a_2 + a_3}\right) \end{cases} \qquad (7.13a)$$

$$E(n_3) = n - E(n_1) - E(n_2) .$$

We can obtain an approximation for $E(n_j^2)$ in a similar way. These approximations for $E(n_j)$ and $E(n_j^2)$ could be used in conjunction with (7.12) or with the less exact approximation

$$E\left(\frac{1}{n_j}\right) \doteq [E(n_j)]^{-1} \qquad (7.13b)$$

To illustrate how to obtain the optimal values of n' and n, we consider the case $D = 2$. Here, the exact expressions for $E(n_1^{-1})$ and $E(n_2^{-1})$ are given by

$$E(\frac{1}{n_1}) = \frac{1}{n_1^0} \sum_{x=n_1^0}^{n_1^0+(n'-n)} P(x) + \sum_{x=1}^{n_1^0-1} \frac{1}{x} P(x)$$

$$+ \sum_{x=n_1^0+(n'-n)+1}^{n'-1} [\frac{1}{x-(n'-n)}] P(x) \qquad (7.14a)$$

and

$$E(\frac{1}{n_2}) = E(\frac{1}{n-n_1}) = (\frac{1}{n-n_1^0}) \sum_{x=n_1^0}^{n_1^0+(n'-n)} P(x)$$

$$+ \sum_{x=1}^{n_1^0-1} [\frac{1}{n-x}]P(x) + \sum_{x=n_1^0+(n'-n)+1}^{n'-1} [\frac{1}{n'-x}]P(x) \qquad (7.14b)$$

where $P(X=x) = \binom{n'}{x}(1-\pi_1)^{n'-x}/\{1-(1-\pi_1)^{n'}-\pi_1^{n'}\}$. Thus, the exact values of $E(n_1^{-1})$ and $E(n_2^{-1})$ can be obtained by using binomial tables.

Several numerical examples are considered in Tables 7.2(a) and 7.2(b). For each one, values of a_1, a_2, π_1, C^*, c' and c are

Table 7.2(a). Values of $E(a_1/n_1)$ and $E(a_2/n_2)$ for Choices of n' and n

n'	n	a_1/a_2	$\{E(a_1/n_1), E(a_2/n_2)\}$		
			$\pi_1 = 0.1$	$\pi_1 = 0.2$	$\pi_1 = 0.4$
45	3	1	0.50, 0.25	0.50, 0.25	0.50, 0.25
		0.6	0.38, 0.31	0.38, 0.31	0.38, 0.31
		1.5	0.31, 0.39	0.30, 0.39	0.30, 0.40
40	4	1	0.27, 0.25	0.25, 0.25	0.25, 0.25
		0.6	0.20, 0.31	0.19, 0.31	0.19, 0.31
		1.5	0.32, 0.20	0.30, 0.20	0.30, 0.20
35	5	1	0.27, 0.16	0.25, 0.17	0.25, 0.17
		0.6	0.20, 0.34	0.19, 0.21	0.19, 0.31
		1.5	0.26, 0.18	0.20, 0.20	0.20, 0.20
30	6	1	0.24, 0.15	0.17, 0.16	0.17, 0.17
		0.6	0.21, 0.15	0.19, 0.16	0.19, 0.16
		1.5	0.26, 0.14	0.16, 0.19	0.15, 0.20
25	7	1	0.25, 0.11	0.18, 0.12	0.17, 0.12
		0.6	0.19, 0.14	0.14, 0.15	0.12, 0.16
		1.5	0.32, 0.13	0.20, 0.14	0.15, 0.13
20	8	1	0.29, 0.09	0.17, 0.11	0.13, 0.12
		0.6	0.22, 0.11	0.15, 0.12	0.13, 0.12
		1.5	0.33, 0.07	0.20, 0.10	0.12, 0.13

Table 7.2(a) (Continued).

n'	n	a_1/a_2	$\{E(a_1/n_1), E(a_2/n_2)\}$		
			$\pi_1 = 0.1$	$\pi_1 = 0.2$	$\pi_1 = 0.4$
15	9	1	0.34, 0.07	0.22, 0.08	0.13, 0.10
		0.6	0.25, 0.09	0.08, 0.10	0.13, 0.11
		1.5	0.40, 0.06	0.26, 0.07	0.13, 0.09
10	10	1	0.38, 0.06	0.29, 0.07	0.15, 0.09
		0.6	0.29, 0.08	0.22, 0.08	0.11, 0.11
		1.5	0.47, 0.05	0.35, 0.05	0.18, 0.07

Table 7.2(b). Values of $E(a_1/n_1)$ and $E(a_2/n_2)$ for Choices of n' and n

n'	n	a_1/a_2	$\{E(a_1/n_1), E(a_2/n_2)\}$		
			$\pi_1 = 0.1$	$\pi_1 = 0.2$	$\pi_1 = 0.4$
34	8	1	0.20, 0.11	0.13, 0.13	0.13, 0.12
		0.25	0.11, 0.14	0.10, 0.13	0.10, 0.13
32	9	1	0.21, 0.09	0.13, 0.10	0.13, 0.10
		0.25	0.11, 0.11	0.10, 0.11	0.10, 0.11
30	10	1	0.22, 0.07	0.12, 0.09	0.10, 0.10
		0.25	0.11, 0.10	0.10, 0.10	0.10, 0.10
28	11	1	0.23, 0.06	0.12, 0.08	0.10, 0.08
		0.25	0.12, 0.09	0.10, 0.09	0.10, 0.09
26	12	1	0.24, 0.06	0.12, 0.07	0.08, 0.08
		0.25	0.11, 0.08	0.10, 0.08	0.10, 0.08
24	13	1	0.26, 0.05	0.13, 0.06	0.09, 0.07
		0.25	0.13, 0.07	0.07, 0.09	0.07, 0.08
22	14	1	0.27, 0.04	0.14, 0.05	0.08, 0.07
		0.25	0.11, 0.07	0.08, 0.07	0.07, 0.07

specified. Then, for each example, the values of $E(a_1/n_1)$ and $E(a_2/n_2)$ are given for various choices of n' and n satisfying the budget constraint. In Table 7.2(a), $C^* = 60$, $c' = 1$ and $c = 5$; in Table 7.2b, $C^* = 50$, $c' = 1$ and $c = 2$. In both tables, $a_1 + a_2 = 1$.

In each case, we can obtain from the tables the optimal values of (n', n). For example, in Table 7.2(a), for $a_1/a_2 = 1$ and $\pi_1 = 0.02$, $(n', n) \doteq (20, 8)$ are the optimal values of (n', n). We may note that in both Tables 7.2(a) and 7.2(b) even moderately large departures of (n', n) from their optimal values do not, in general, result in large increases in $\max\{a_1/n_1 , a_2/n_2\}$. We note that for Table 7.2(a), we considered all pairs of (n', n) satisfying the budget constraint. However, to simplify our computations, in Table 7.2(b), we used the crude approximations (7.13a) and (7.13b) to obtain the "approximate" value of optimal (n', n). The approximate value thus obtained for $\pi_1 = 0.2$ and $a_1/a_2 = 1$ was $(n', n) \doteq (26, 12)$. Then we obtained the exact values of (n', n) in the neighborhood of the "approximate" optimal values of (n', n).

VIII. BIBLIOGRAPHY

[1] Aggarwal, Om P. Bayes and minimax procedures in sampling from finite and infinite populations - I. Annals of Mathematical Statistics 30: 206-218. 1959.

[2] Aoyama, Hirojiro. A study of stratified random sampling. Annals of the Institute of Statistical Mathematics (Tokyo, Japan) 6(1): 1-36. 1954.

[3] Bose, C. Note on the sampling error in the method of double sampling. Sankhya 6: 330. 1943.

[4] Chaddha, R. L., Hardgrave, W. W., Hudson, D. J., Segal, M., Suurballe, J. W. and Tischendorf, J. A. Allocation of total sample size when the stratum means are of interest. Unpublished mimeographed manuscript. Holmdel, N.J., Bell Telephone Laboratories, Inc. 1969.

[5] Chakravarti, I. M., Laha, R. G. and Roy, J. Handbook of methods of applied statistics. Vol. II: Planning of surveys and experiments. New York, N.Y., John Wiley and Sons, Inc. 1967.

[6] Chikkagoudar, M. S. On pps sampling with and without replacement. Australian Journal of Statistics 9(3): 109-118. 1967.

[7] Chikkagoudar, M. S. Two-phase sampling for pps estimation. Annals of the Institute of Statistical Mathematics (Tokyo, Japan) 19(1): 133-142. 1967.

[8] Cochran, W. G. Comparison of methods for determining stratum boundaries. Bulletin De L'Institut International de Statistique 38(2): 345-358. 1961.

[9] Cochran, W. G. Sampling techniques. 2nd ed. New York, N.Y., John Wiley and Sons, Inc. 1963.

[10] Dalenius, T. Sampling in Sweden. Stockholm, Sweden, Almquist and Wicksell. 1957.

[11] Dalenius, T. and Hodges, Joseph L. Choice of stratification points. Skandinavisk Aktuarietidskrift 40: 198-203. 1957.

[12] Dalenius, T. and Hodges, J. L. Minimum variance stratification. Journal of the American Statistical Association 54: 88-101. 1959.

[13] Durbin, J. A note on the application of Quenouille's method of bias reduction to the estimation of ratios. Biometrika 46: 477-480. 1959.

[14] Ekman, G. An approximation useful in univariate stratification. Annals of Mathematical Statistics 30: 219-229. 1959.

[15] Ekman, G. Approximate expressions for the conditional mean and variance over small intervals of a continuous distribution. Annals of Mathematical Statistics 30: 1131-1134. 1959.

[16] Frauendorfer, R. Numerical analysis of some unbiased and approximately unbiased ratio-type estimators. Unpublished M. S. thesis. Ames, Iowa, Library, Iowa State University. 1967.

[17] Fuller, W. A. and Johnson, N. R. On the bias and MSE of some regression estimators. Unpublished mimeographed manuscript. Iowa Agriculture and Home Economics Experiment Station No. J-6005. 1969.

[18] Ghosh, S. P. Post-cluster sampling. Annals of Mathematical Statistics 34: 587-597. 1963.

[19] Goodman, Leo A. On the exact variance of products. Journal of the American Statistical Association 55: 708-713. 1960.

[20] Goodman, Leo A. and Hartley, H. O. The precision of unbiased ratio-type estimators. Journal of the American Statistical Association 53: 491-508. 1958.

[21] Goswami, J. N. and Sukhatme, B. V. Ratio method of estimation in multi-phase sampling with several auxilliary variables. Indian Society of Agricultural Statistics 17: 83-103. 1965.

[22] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. Sample survey methods and theory Vol. I: Methods and applications. New York, N. Y., John Wiley and Sons, Inc. 1953.

[23] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. Sample survey methods and theory. Vol. II: Theory. New York, N. Y., John Wiley and Sons, Inc. 1953.

[24] Hartley, H. O. and Rao, J. N. K. Sampling with unequal probabilities and without replacement. Annals of Mathematical Statistics 33: 350-374. 1962.

[25] Hartley, H. O. and Ross, A. Unbiased ratio estimates. Nature 174: 270-271. 1954.

[26] Herlekar, R. K. The problem of optimum stratification. Part I. Investigation into some two stage stratified sampling procedures for populations represented by probability density functions. Skandinavisk Aktuarietidiskrift 50: 1-18. 1967.

[27] Herlekar, R. K. The problem of optimum stratification. Part II. An investigation into a two-stage sampling procedure applied to a bivariate normal population. Skandinavisk Aktuarietidskrift 50: 183-202. 1967.

[28] Jumbunathan, M. V. A note on the efficiency of double sampling for stratification. Sankhya 22: 367-370. 1960.

[29] Kendall, M. G. and Stuart, A. The advanced theory of statistics Vol. I: Distribution theory. 2nd ed. London, England, Charles Griffin and Company, Ltd. 1963.

[30] Kish, Leslie. Survey sampling. New York, N. Y., John Wiley and Sons, Inc. 1967.

[31] Kono, K. Note on the double sampling method. Bulletin of Mathematical Statistics, Research Association of Statistical Sciences (Fukuoka, Japan) 4, Nos. 1-2: 36-38. 1950.

[32] Koop, J. C. A note on the bias of the ratio estimator. Bulletin of the International Statistical Institute 33(2): 141-146. 1951.

[33] Lahiri, D. B. A method of sample selection providing unbiased ratio and regression estimators. Bulletin of the International Statistical Institute 33(2): 133-140. 1951.

[34] Mickey, M. R. Some finite population unbiased ratio and regression estimators. Journal of the American Statistical Association 54: 594-612. 1959.

[35] Murakami, M. Some considerations on the ratio and regression estimates. Bulletin of Mathematical Statistics, Research Association of Statistical Sciences (Fukuoka, Japan) 4: 39-42. 1950.

[36] Neyman, J. Contribution to the theory of sampling human populations. Journal of the American Statistical Association 33: 101-116. 1938.

[37] Pascual, Jose Nieto de. Unbiased ratio estimators in stratified sampling. Journal of the American Statistical Association 56: 70-87. 1961.

[38] Raj, Des. On double sampling for pps estimation. Annals of Mathematical Statistics 35: 900-902. 1964.

[39] Raj, Des. On forming strata of equal aggregate size. Journal of the American Statistical Association 59: 481-486. 1964.

[40] Raj, Des. Sampling theory. New York, N.Y., McGraw-Hill Book Company. 1968.

[41] Rao, J. N. K. A note on the estimation of ratios by Quenouille's method. Biometrika 52: 647-649. 1965.

[42] Rao, J. N. K. The precision of Mickey's unbiased ratio estimator. Biometrika 54: 321-324. 1967.

[43] Rao, J. N. K. Ratio and regression estimators. Invited paper for the symposium on the Foundations of survey sampling, University of North Carolina, Chapel Hill. College Station, Texas, Texas A and M. 1968.

[44] Rao, J. N. K. and Beegle, LeNelle D. A Monte Carlo study of some ratio estimators. Social Statistics Section, American Statistical Association Proc. 1966: 443-449. 1966.

[45] Rao, J. N. K., Hartley, H. O. and Cochran, W. G. On a simple procedure of unequal probability sampling without replacement. Journal of Royal Statistical Society, Series B, 24: 482-491. 1962.

[46] Rao, J. N. K. and Webster, J. T. On two methods of bias reduction in the estimation of ratios. Biometrika 53: 571-577. 1966.

[47] Rao, T. J. On certain unbiased ratio estimators. Annals of the Institute of Statistical Mathematics (Tokyo) 18: 117-121. 1966.

[48] Rao, T. J. On the variance of the ratio estimator for Midzuno-Sen sampling scheme. Metrika 10: 89-91. 1966.

[49] Sedransk, J. A double sampling scheme for analytical surveys. Journal of the American Statistical Association 60: 985-1004. 1965.

[50] Serfling, R. J. Approximately optimal stratification. Journal of the American Statistical Association: 63: 1298-1309. 1968.

[51] Singh, D. and Singh, B. D. Some contributions to two-phase sampling. The Australian Journal of Statistics 7(2): 45-47. 1965.

[52] Singh, M. P. The relative efficiency of some two-phase sampling schemes. Annals of Mathematical Statistics 38: 937-940. 1967.

[53] Srivastava, S. R. An estimator using auxiliary information in sampling surveys. Calcutta Statistical Association Bulletin 16: 120-132. 1967.

[54] Sukhatme, B. V. Some ratio-type estimators in two-phase sampling. Journal of the American Statistical Association 57: 628-632. 1962.

[55] Sukhatme, B. V. and Koshal, R. S. A contribution to double sampling. Journal of the Indian Society of Agricultural Statistics 11: 128-144. 1959.

[56] Sukhatme, P. V. Sampling theory of surveys with applications. Unpublished dittoed revised text. Ames, Iowa, Statistical Laboratory, Iowa State University. 1968.

[57] Tikkiwal, B. D. On the theory of Classical regression and double sampling estimation. Journal of the Royal Statistical Society Series B, 22: 131-138. 1960.

[58] Tin, M. Comparison of some ratio estimators. Journal of the American Statistical Association 60: 294-307. 1965.

[59] United Nations. Manual on Sampling Theory. New York, N. Y., United Nations Publications. 1960.

[60] Williams, W. H. Generating unbiased ratio and regression estimators. Biometrics 17: 267-274. 1961.

[61] Williams, W. H. On two methods of unbiased estimation with auxiliary variates. Journal of the American Statistical Association 57: 184-186. 1962.

[62] Williams, W. H. Post-stratification estimates for multistage samples. Unpublished M.S. thesis. Ames, Iowa, Library, Iowa State University. 1956.

[63] Yates, F. Sampling methods for censuses and surveys. 3rd ed. London, England, Charles Griffin and Company, Ltd. 1960.

IX. ACKNOWLEDGMENTS

I take this opportunity to thank Dr. J. Sedransk for his generous guidance and encouragement during the preparation of this dissertation. My thanks go also to my wife, Beatrice, to Dr. B. Gil, formerly United Nations Census expert in Ghana, to Dr. J. C. Caldwell of the Population Council of America and to others who made this study possible. Thanks are given also to Eric West who wrote the computer programs for the numerical illustrations in Chapter V.

The research was supported mainly by the Population Council of America and the Government of Ghana. Some support was also given by the U. S. Office of Education under contract number OEC-3-6-002041-2041. The above support is gratefully acknowledged.